29TH INTERNATIONAL CONFERENCE ON COMPUTERS IN EDUCATION

ICCE 2021

November 22-26, 2021

CONFERENCE PROCEEDINGS

Volume I

29th International Conference on Computers in Education Conference Proceedings Volume I

Editors: Maria Mercedes T. RODRIGO, Sridhar IYER, Antonija MITROVIC

Publisher: Asia-Pacific Society for Computers in Education (APSCE)
Address: Center for Science and Technology for Learning,
National Central University,No. 300, Jhongda Road, Zongli District,
Taoyuan City 32001, Taiwan
Telephone: +886-3-4227151 ext. 35407
FAX: +886-3-4227151 ext. 35407
Email: service@apsce.net
Website: http://www.apsce.net

Date of publication: November 22, 2021

ISBN: ISBN 978-986-97214-7-9 (PDF)

EDITORS

Maria Mercedes T. RODRIGO, Ateneo de Manila University, Philippines
Sridhar IYER, Indian Institute of Technology, India
Antonija MITROVIC, University of Canterbury, New Zealand

ASSOCIATE EDITORS

(in alphabetical order)

MESSAGE FROM THE CONFERENCE CHAIRS

Antonija MITROVIC

Conference Chair
University of Canterbury,
New Zealand

On behalf of the organizing committee, I would like to welcome all participants of the 29th International Conference on Computers in Education (ICCE) 2021, the flagship conference series of the Asia-Pacific Society for Computers in Education (APSCE). Last year, we had the first virtual conference, which went well, but we were all hoping that ICCE 2021 would be face-to-face. Unfortunately, that is still not possible. We planned to meet in Bangkok, Thailand, and enjoy the hospitality of our Local organizing committee: Thepchai Supnithi, Niwat Srisawasi and Charoenchai Wongwatkit. We thank them and their team for all their efforts on organizing the conference in Bangkok. They are preparing some virtual content for all of us to make the conference more enjoyable.

The pandemic has brought a lot of anxiety, uncertainty and changes. I would like to thank our standing committee, for being flexible and finding solutions to challenging problems we faced. Our appreciation goes to Pham-Duc Tho, the Managing Secretary of APSCE, the tech wizard who made ICCE 2020 possible. He is still continuing to work tirelessly for APSCE and the conference.

My sincere appreciation goes to Didith Rodrigo and Sridhar Iyer, the chair and co-chair of the International Program Committee respectively. They have put an enormous amount of time in making sure that we have excellent programme at ICCE 2021. My gratitude goes to the chairs of the seven subconferences, organizers of workshops, tutorials, panels, WIPP, DSC, ES, posters, and ECW. And of course, our sincere thanks to all authors, reviewers, presenters, Doctoral students, and other participants. I would also like to thank our consultants, Lung-Hsiang Wong, Hyo-Jeong So and Jon Mason, for sharing their wisdom and advising us along the way.

Four outstanding keynote speakers will share their insights across varying areas in the field of computers in education. The first speaker is Kulthida Tuamsuk, from the Khon Kaen University, Thailand, and she will present at talk about transforming classrooms into learning communities at her university. Pierre Dillenbourg from the Swiss Federal Institute of Technology will talk about his experience with orchestrating classrooms. Tiffany Barnes, from the North Carolina University, will talk about using both human and artificial intelligence to provide better learnig experiences. Gwo-Jen Hwang from the National Taiwan University of Science and Technology will talk about issues in using AIED in the mobile era.

There will also be three equally inspiring theme-based invited speeches. Ana Gimeno from the Universitat Politechnica de Valencia, Spain, will talk about whether MOOCs satisfy learner needs. Baltasar Fernández-Manjón, from the Complutense University of Madrid, Spain, will talk about game learning analytics Jon Mason from the Charles Darwin University, Australia, will talk about questioning in the digital environment.

I hope the participants will find the conference invigorating, relevant and enjoyable!

MESSAGE FROM THE INTERNATIONAL PROGRAM COORDINATION CHAIRS

Maria Mercedes T. RODRIGO

International Program Coordination Chair
Ateneo de Manila University, Philippines

Sridhar IYER

International Program Coordination Co-Chair
Indian Institute of Technology, India

Welcome to the 29th International Conference on Computers in Education (ICCE)! Organized by the Asia-Pacific Society for Computers in Education, this annual conference is a venue in which scholars, researchers, and academics share their work regarding the use of Information and Communication Technology (ICT) in various settings of education.

Because of the continuing COVID-19 pandemic and the travel difficulties related to stemming its spread ICCE 2021 is held virtually from November 22 to November 26, 2021. As with last year's conference, we believe gathering in a virtual space keeps us safer while still allowing us to engage in meaningful and productive dialogs.

ICCE 2021 continues the meta-conference tradition of the previous ICCEs. As such, the conference is organized into seven sub-conference programs specializing specific themes:

- C1: ICCE Sub-Conference on Artificial Intelligence in Education/Intelligent Tutoring System (AIED/ITS)
- C2: ICCE Sub-Conference on Computer-supported Collaborative Learning (CSCL) and Learning Sciences (LS)
- C3: ICCE Sub-Conference on Advanced Learning Technologies (ALT), Learning Analytics, Platforms and Infrastructure
- C4: ICCE Sub-Conference on Classroom, Ubiquitous, and Mobile Technologies Enhanced Learning (CUMTEL)
- C5: ICCE Sub-Conference on Educational Gamification and Game-based Learning (EGG)
- C6: ICCE Sub-Conference on Technology Enhanced Language Learning (TELL)
- C7: ICCE Sub-Conference on Practice-driven Research, Teacher Professional Development and Policy of ICT in Education (PTP)

The International Program Committee is led by a strong and dedicated team, which includes the Conference Chair, the Program Coordination Chair and Co-Chair, 52 executive Sub-Conference Chairs and Co-Chairs and 253 experts in the field of Computers in Education from 27 different countries or economies. Former ICCE local organizing and program coordination chairs have played important roles as consultants in overseeing the organization process of this conference.

ICCE 2020 received a total of 148 submissions (116 full, 28 short, and 4 posters) from 25 different countries or economies. Top three countries with the highest number of submissions are Japan, China, and the Philippines. Submissions were also received from the Middle East, Europe, America and Africa, which signals the international interest toward ICCE 2020. Table 1 provides the submissions statistics by the country of the first author:

Table 1. Submission statistics by country (based on first author's country)

Countries or Economy			
Australia	4	New Zealand	2
Bangladesh	1	Norway	1
Canada	1	Philippines	18
China	18	Poland	1
Egypt	1	Singapore	3
Germany	1	Spain	4
Hong Kong	6	Sweden	1
India	16	Taiwan	10
Indonesia	1	Thailand	10
Iran	1	Tunisia	3
Israel	1	Turkey	3
Japan	33	United States	6
Malaysia	2		
Total			**148**

All papers were subjected to a rigorous review process by at least three reviewers from the respective Sub-Conference program committees. After the reviews were completed, a meta-review was provided for each paper. In total, 650 reviews and meta-reviews were received. After the discussion period within the individual program committees led by the Sub-Conference Executive Chairs and Co-Chairs, recommendations were made to the Program Coordination Committee Chair and Co-Chair, who oversaw the review process and quality for all Sub-Conferences. This resulted in 30 full papers, 57 short papers, and 22 posters accepted across seven Sub-Conferences. The overall acceptance rate for full papers is 25.9%, which reflects our efforts to continue the maintenance of the quality of presentations at ICCE 2021. The complete statistics of paper acceptance is shown in Table 2.

Table 2. Paper Acceptance Statistics

	Submission	**Submit as Full**	**Accepted as Full**	**Full %**	**Accepted as Short**	**Accepted as Poster**	**Overall %**
C1 - AIED/ITS	26	21	5	23.8%	9	4	69.2%
C2 - CSCL/LS	16	10	2	20.0%	4	4	62.5%
C3 - ALT/LA	32	26	8	30.8%	10	7	78.1%
C4 - CUMTEL	17	14	1	7.1%	11	1	76.5%
C5 - EGG	14	10	3	30.0%	6	1	71.4%
C6 - TELL	16	13	3	23.1%	7	2	75.0%
C7 - PTP	27	22	8	36.4%	10	3	77.8%
Total	**148**	**116**	**30**	**25.9%**	**57**	**22**	73.6%

In addition to the main program with seven sub-conferences, ICCE 2021 includes various program components, such as Keynote Speeches, Theme-based Invited Speeches, Workshops, Tutorials, Work-in-Progress Posters (WIPP), Extended Summary (ES), Doctoral Student Consortia (DSC), and Early Career Workshop (ECW). All the papers in these program components are compiled and published in a separate volume with its own ISBN. Pre-conference events are held on the first two days of the conference, including 12 workshops, one tutorial, three panels, DSC, ECW, APSCE Student Wing Workshop, and SIG community building sessions.

We are grateful to all who contributed to ICCE 202's success. We thank all the paper authors for choosing ICCE 2021 as the venue to present their research. We would also like to thank the IPC Executive Chairs/Co-Chairs and members, who undertook the responsibility of reviewing and selecting papers that represent research of high quality. Specially thanks to our Keynote and Invited Speakers for accepting our invitations and sharing inspiring research with the ICCE 2021 participants.

In the challenging times we face and those that continue to come, our work becomes more relevant and important than ever. We are grateful to the APSCE community for staying focused and resilient, and for continuing to make the valuable contributions to education that will help shape the minds, hearts, and spirits of future generations.

MESSAGE FROM THE LOCAL ORGANIZING COMMITTEE CHAIRS

Thepchai SUPNITHI

National Electronics and Computer Technology Center, Thailand

Niwat SRISAWASDI

Faculty of Education, Khon Kaen University, Thailand

Charoenchai WONGWATKIT

Mae Fah Luang University, Thailand

Welcome, all the researchers and participants around the globe to Thailand virtually in ICCE 2021.

This year, our Local Organizing Committee Chairs (LOC) from National Electronics and Computer Technology Center (NECTEC), Khon Kaen University (KKU), and Mae Fah Luang University (MFU), Thailand are greatly privileged to host The 29th International Conference on Computers in Education (ICCE 2021). Thailand is one of the best locations globally, with various activities, food, culture, and travel destinations. In academia, Thailand always successfully hosts and welcomes many scholars, researchers, and professionals through various international academic events.

ICCE 2021 celebrates turning to the third decade of this successfully well-structured event of computers in education. With COVID-19 outbreak, traveling is highly restricted to many countries. ICCE 2021 is presented as a fully virtual conference featuring seven main conferences, twelve pre-conference workshops, an interactive tutorial, early career workshops, doctoral student consortium, work-in-progress events, keynote speakers, and theme-based invited speakers.

We would like to take this opportunity to acknowledge strong partnerships with The Asia-Pacific Society for Computers in Education (APSCE) in making this conference successful. Sincere gratitudes go to the contributions from the international organizing committee, authors, participants, supporters, and sponsors inevitably. With solid collaborations, ICCE 2021 has become one of the amazing conference events in this area.

We believe that this conference is full of wonderful moments, fruitful discussions, and beautiful community buildings.

See you in ICCE 2021. Thank you! ขอบคุณครับ

ORGANIZATION

Organized by: Asia Pacific Society for Computers in Education

Standing Committee

- *Conference Chair*
  Antonija MITROVIC, University of Canterbury, New Zealand
- *International Program Coordination Chair*
  Maria Mercedes T. RODRIGO, Ateneo de Manila University, Philippines
- *International Program Coordination Co-Chair*
  Sridhar IYER, Indian Institute of Technology, India
- *Local Organizing Committee Chairs*
  Thepchai SUPNITHI, National Electronics and Computer Technology Center,Thailand
  Niwat SRISAWASDI, Khon Kaen University, Thailand
  Charoenchai WONGWATKIT, Mae Fah Luang University, Thailand
- *Consultants*
  Hyo-Jeong SO, Eawha Womans University, South Korea
  Jon MASON, Charles Darwin University, Australia

Sub-Conference
C1: Artificial Intelligence in Education/Intelligent Tutoring System (AIED/ITS) and Adaptive Learning

- *PC Executive Chair*
  Kazuhisa SETA, Osaka Prefecture University, Japan
- *PC Co-Chair*
  Patcharin PANJABUREE, Mahidol University, Thailand
  Nguyen-Thinh LE, Humboldt-Universität zu Berlin, Germany
  Valéry PSYCHÉ, Université TÉLUQ, Canada

C2: Computer-supported Collaborative Learning (CSCL) and Learning Sciences

- *PC Executive Chair*
  Camillia MATUK, New York University, USA
- *PC Executive Co-Chair*
  Kate THOMPSON, Queensland University of Technology, Australia
  Daniel BODEMER, University of Duisburg-Essen, Germany
  Sahana MURTHY, Indian Institute of Technology Bombay, India
  Elizabeth Ruilin KOH, National Institute of Education, Singapore

C3: Advanced Learning Technologies (ALT), Learning Analytics and Digital Infrastructure

- *PC Executive Chair*
  Ramkumar RAJENDRAN, Indian Institute of Technology Bombay, India
- *PC Executive Co-Chair*
  Rwitajit MAJUMDAR, Kyoto University, Japan
  Khalid KHAN, Charles Darwin University, Australia
  Ismar Frango SILVEIRA, Mackenzie Presbyterian University, Brazil
  Mohammed SAQR, University of Eastern Finland, Finland

Aditi KOTHIYAL, EPFL, Switzerland

C4: Classroom, Ubiquitous, and Mobile Technologies Enhanced Learning (CUMTEL)

- *PC Executive Chair*
  Jingyun WANG, Durham University, UK
- *PC Executive Co-Chair*
  Michelle BANAWAN, Arizona State University, USA
  Brendan FLANAGAN, Kyoto University, Japan
  Andrea VALENTE, University of Southern Denmark, Denmark

C5: Educational Gamification and Game-based Learning (EGG)

- *PC Executive Chair*
  Hercy N. H. CHENG, Central China Normal University, China
- *PC Executive Co-Chair*
  Borja MANERO, Universidad Complutense de Madrid, Spain
  Mouna DENDEN, University of Tunis, Tunisia
  Armando Maciel TODA, University of São Paulo

C6: Technology Enhanced Language Learning (TELL)

- *PC Executive Chair*
  Agnieszka PALALAS, Athabasca University, Canada
- *PC Executive Co-Chair*
  Mark PEGRUM, University of Western Australia, Australia
  Weichao "Vera" CHEN, Baylor College of Medicine, USA
  Olga VIBERG, KTH Royal Institute of Technology, Sweden

C7: Practice-driven Research, Teacher Professional Development and Policy of ICT in Education (PTP)

- *PC Executive Chair*
  Dan KOHEN-VACS, Holon Institute of Technology, Israel
- *PC Executive Co-Chair*
  Gautam BISWAS, Vanderbilt University, USA
  Marc JANSEN, HOCHSCHULE RUHR WEST, Germany
  Mishra SHITANSHU, Indian Institute of Technology Bombay, India
  Ivica BOTICKI, University of Zagreb, Croatia

Workshop & Interactive Events

- *Chair*
  Charoenchai WONGWATKIT, Mae Fah Luang University, Thailand
- *Co-Chair*
  Chiu-Lin LAI, National Taipei University of Education, Taiwan
  Sasithorn CHOOKAEW, King Mongkut's University of Technology North Bangkok, Thailand

Tutorials

- *Chair*
  Rustam SHADIEV, Nanjing Normal University, China

- *Co-Chair*
  Kaushal Kumar BHAGAT, G. S. Sanyal School of Telecommunications, India

Work-in-Progress Poster (WIPP)

- *Chair*
  Kazuaki KOJIMA, Teikyo University, Japan
- *Co-Chair*
  Mi Song KIM, University of Western Ontario, Canada

Doctoral Student Consortium (DSC)

- *Chair*
  Morris JONG, The Chinese University Hong Kong, Hong Kong
- Co-Chair
  Hiroaki OGATA, Kyoto University, Japan

Early Career Workshop (ECW)

- *Chair*
  Mas Nida BT MD KHAMBARI, Universiti Putra Malaysia, Malaysia
- *Co-Chair*
  Ryan EBARDO, Jose Rizal University, Philippines

Panels

- *Chair*
  Han-Yu SUNG, National Taipei University of Nursing and Health Science, Taiwan
- *Co-Chair*

Extended Summary (ES)

- *Chair*
  Ping LI, The Hong Kong Polytechnic University, Hong Kong
- *Co-Chair*

Merit Scholarships

- *Chair*
  Mohammad Nehal HASNINE, Hosei University, Japan
- *Co-Chair*
  Gökhan AKÇAPINAR, Hacettepe University, Turkey

Special Interest Groups (SIG)

- ***S1: Artificial Intelligence in Education/Intelligent Tutoring Systems/Adaptive Learning (AIED/ITS/AL)***
  Michelle P. BANAWAN, Arizona State University, USA
- ***S2: Computer-supported Collaborative Learning and Learning Sciences (CSCL)***
  Chew Lee TEO, Nanyang Technological University, Singapore
- ***S3: Advanced Learning Technologies, Learning Analytics, Platforms and Infrastructure (ALT)***
  Jin Gon SON, Korea National Open University, Korea
- ***S4: Classroom, Ubiquitous, and Mobile Technologies Enhanced Learning (CUMTEL)***

Ting-Chia HSU, National Taiwan Normal University, Taiwan

- ***S5: Educational Gamification and Game-based Learning (EGG)***
  Rita KUO, New Mexico Institute of Mining and Technology, Taiwan
- ***S6: Technology Enhanced Language Learning (TELL)***
  Yoshiko GODA, Kumamoto University, Japan
- ***S7: Practice-driven Research, Teacher Professional Development and Policy of ICT in Education (PTP)***
  Sahana MURTHY, Indian Institute of Technology Bombay, India
- ***S8: Development of Information and Communication Technology in the Asia-Pacific Neighborhood (DICTAP)***
  Bo JIANG, Zhejiang University of Technology, China
- ***S9: Educational Use of Problems/Questions in Technology-Enhanced Learning (EUPQ)***
  Kazuaki KOJIMA, Teikyo University, Japan
- ***S10: Learning Analytics and Educational Data Mining (LAEDM)***
  Brendan FLANAGAN, Kyoto University, Japan
- ***S11: Computational Thinking Education & STEM Education (CTE&STEM)***
  Siu Cheung KONG, The Education University of Hong Kong, Hong Kong

C1 PC Members

- Kazuhisa Seta, Osaka Prefecture University, Japan
- Patcharin PanjabureE, Mahidol University, Thailand
- Nguyen-Thinh Le, Humboldt-Universität zu Berlin, Germany
- Valéry Psyché, Université TÉLUQ, Canada
- Sagaya Amalathas, Taylors University, Malaysia
- Ryan Baker, University of Pennsylvania, United States
- Kritya Bunchongchit, Mahidol University, Thailand
- Chih-Yueh Chou, Yuan Ze University, Taiwan
- Philippe Fournier-Viger, Harbin Institute of Technology, China
- Claude Frasson, University of Montreal, Canada
- Yuki Hayashi, Osaka Prefecture University, Japan
- Bastiaan Heeren, Open University, The Netherlands
- Tomoya Horiuchi, Kobe University, Japan
- Sharon Hsiao, Arizona State University, United States
- Akihiro Kashihara, The University of Electro-Communications, Japan
- Tomoko Kojiri, Kansai University, Japan
- Tatsuhiro Konishi, Shizuoka University, Japan
- Noboru Matsuda, North Carolina State University, United States
- Tatsunori Matsui, Waseda University, Japan
- Antonija Mitrovic, University of Canterbury, New Zealand
- Riichiro Mizoguchi, Japan Advanced Institute of Science and Technology, Japan
- Roger Nkambou, Université du Québec à Montréal, Canada
- Ange Tato, Université du Québec à Montréal, Canada
- Jose Luis Perez De La Cruz, Universidad de Malaga, Spain
- Elvira Popescu, University of Craiova, Romania
- Maria Mercedes T. Rodrigo, Ateneo de Manila University, Philippines

- Olga C.Santos, aDeNu Research Group, Spain
- John Stamper, Carnegie Mellon University, United States
- Thepchai Supnithi, NECTEC, Thailand
- Benedict Du Boulay, University of Sussex, United Kingdom

C2 PC Members

- Sahana Murthy, Indian Institute of Technology Bombay, India
- Camillia Matuk, New York University, United States
- Bodong Chen, University of Minnesota, United States
- Elizabeth Ruilin Koh, National Institute of Education, Singapore
- Sven Manske, University of Duisburg-Essen, Germany
- Kate Thompson, Queensland University of Technology, Australia
- Daniel Bodemer, University of Duisburg-Essen, Germany
- Rose Liang, National University of Singapore, Singapore
- Mike Tissenbaum, University of Illinois at Urbana-Champaign, United States
- Amanda Dickes, Gulf of Maine Research Institute, , United States
- Simon Leonard, University of South Australia, Australia
- Cynthia D'Angelo, University of Illinois at Urbana-Champaign, United States
- Brendan Eagan, University of Wisconsin-Madison, United States
- Kristin Searle, Utah State University, United States
- Steven Kickbusch, Queensland University of Technology, Australia
- Hongzhi Yang, The University of Sydney, Australia
- Marcelo Worsley, Northwestern University, United States
- Lenka Schnaubert, University of Duisburg-Essen, Germany
- Tamara Galoyan, University of Utah, United States
- Nick Kelly, Queensland University of Technology, Australia
- Sanjay Chandrasekharan, Homi Bhabha Centre for Science Education, India
- Shilpa Sahay, New York University, United States
- Antonette Shibani, University of Technology, Sydney
- Florence Gabriel, University of South Australia, Australia
- Pei-Yi Lin, National Kaohsiung Normal University, Taiwan
- Iwona Czaplinski, Queensland Univeristy of Technology, Australia
- Natasha Arthars, The University of Sydney, Australia
- Julia Eberle, Ruhr-University Bochum, Germany
- Jürgen Buder, Leibniz-Institut für Wissensmedien, Germany
- Simon Knight, University of Technology, Sydney
- Sarah Dart, Queensland University of Technology, Australia
- Johanna Pöysä-Tarhonen, University of Jyvaskyla, Finland
- Seng Chee Tan, Nanyang Technological University, Singapore
- Chandan Dasgupta, Indian Institute of Technology Bombay, India
- Polly Lai, Queensland University of Technology, Australia
- Chew Lee Teo, Ministry of Education, Singapore

C3 PC Members

- Oluwafemi Samson Balogun, University of Eastern Finland, Finland
- Vladimir Costas, Universidad Mayor de San Simón, Bolivia
- Regina Motz, Universidad de la República, Uruguay
- Sonsoles López-Pernas, Universidad Politécnica de Madrid, Spain
- Gökhan Akçapınar, Hacettepe University, Turkey
- Aditi Kothiyal, Swiss Federal Institute of Technology in Lausanne, Switzerland
- Khalid Khan, Charles Darwin University, Australia
- Ramkumar Rajendran, Indian Institute of Technology Bombay, India
- Rwitajit Majumdar, Kyoto University, Japan
- Rekha Ramesh, Mumbai University, India
- Maria Eliseo, Universidade Presbiteriana Mackenzie, Brazil
- Mohammed Saqr, University of Eastern Finland
- Jon Mason, Charles Darwin University, Australia
- Yang-Hsueh Chen, National Chengchi Universitry, Taiwan
- Lenka Schnaubert, University of Duisburg-Essen, Germany
- Atsushi Shimada, Kyushu University, Japan
- Emmanuel Awuni Kolog, University of Ghana Business School, Ghana
- Łukasz Tomczyk, Pedagogical University of Cracow, Poland
- Erkan Er, Middle East Technical University, Turkey
- Yongwu Miao, University Duisburg-Essen, Germany
- Ismar Silveira, Universidade Presbiteriana Mackenzie, Brazil
- Minhong Wang, The University of Hong Kong, Hong Kong
- Judith Azcarraga, De La Salle University, Philippines
- Riina Vuorikari, Institute for Prospective Technological Studies
- Victoria Abou Khalil, Kyoto University, Japan
- Mehmet Kokoç, Karadeniz Technical University, Turkey
- Nigel Stanger, University of Otago, Japan
- Luis Anido Rifon, Universidade de Vigo, Spain
- Manuel Caeiro Rodríguez, University of Vigo, Spain
- Marc Jansen, University of Applied Sciences Ruhr West, Germany
- Tore Hoel, Oslo Metropolitan University, Norway
- Weiqin Chen, Oslo Metropolitan University, Norway
- Chew Lee Teo, Ministry of Education, Singapore
- Jerry Chih-Yuan Sun, National Chiao Tung University, Taiwan

C4 PC Members

- 國 豪 黃, National Yunlin University of Science & Technology, Taiwan
- Tai-Chien Kao, National Dong Hwa University, Taiwan
- Su Cai, Beijing Normal University, China
- Jingyun Wang, Durham University, England
- Brendan Flanagan, Kyoto University, Japan
- Michelle Banawan, Arizona State University, United States
- Longkai Wu, National Institute of Education, Singapore

- Daner Sun, The Education University of Hong Kong, Hong Kong
- Mi Song Kim, Western University, Canada
- Kai-Hsiang Yang, National Taipei University of Education, Taiwan
- Iwen Huang, National University of Tainan, Taiwan
- Ivica Boticki, University of Zagreb, Croatia
- Yih-Ruey Juang, Jinwen University of Science and Technology
- Yiu Chi Lai, The Education University of Hong Kong, Hong Kong
- Bian Wu, East China Normal University, China
- Peter Wan, The Education University of Hong Kong, Hong Kong
- Chengjiu Yin, Kobe University, Japan
- Fuhua Lin, Athabasca University, Canada
- Guang Chen, Beijing Normal University, China
- Chen-Yu Lee, Ling Tung University, Taiwan
- Ping He, Tianjin University, China
- Tzu Chi Yang, National Taipei University of Education, Taiwan
- Martina Holenko Dlab, University of Rijeka, Croatia
- Kuo-Liang Ou, National Tsing Hua University, Taiwan
- Kaushal Kumar, Indian Institute of Technology, India
- Igor Mekterović, University of Zagreb, Croatia
- M. Carmen Juan, Universitat Politècnica de València, Spain
- Andrea Valente, University of Southern Denmark, Denmark
- Haiguang Fang, Capital Normal University, China
- Gwo-Jen Hwang, National Taiwan University of Science and Technology, Taiwan
- Huiying Cai, Jiangnan University, China
- Ting-Ting Wu, National Yunlin University of Science & Technology, Taiwan
- Xuefeng Wei, Ludong University, China
- Chih-Ming Chu, National Ilan University

C5 PC Members

- Armando Maciel Toda, University of São Paulo, Brazil
- Borja Manero, Universidad Complutense de Madrid, Spain
- Demetrios Sampson, Curtin University, Australia
- Achraf Othman, Mada center, Qatar
- Mohamed Koutheaïr Khribi, Mada center, Qatar
- Tzu-Chao Chien, National Central University, Taiwan
- Hercy N. H. Cheng, Taipei Medical University, Taiwan
- Sheng-Kai Yin, Cheng Shiu University, Taiwan
- Michal Ptaszynski, Kitami Institute of Technology, Japan
- Mouna Denden, University of Tunis, Tunisia
- Jorge Simoes, Higher Polytechnic Institute of Gaya, Portugal
- Alexandra I Cristea, Durham University, England
- Zhi-Hong Chen, National Taiwan Normal University, Taiwan
- Shu-Yuan Tao, Takming University of Science and Technology, Taiwan
- Sabine Graf, Athabasca University, Canada

- Ju-Ling Shih, National Central University, Taiwan
- Toshihiro Hayashi, Kagawa University, Japan
- Susan Gwee, English Language Institute of Singapore, Singapore
- Pedro Wightman, Universidad del Norte-Uninorte, Colombia
- Fathi Essalmi, University of Jeddah, Saudi Arabia
- Kaoru Sumi, Future University Hakodate, Japan
- Wing-Kwong Wong, National Yunlin University of Science & Technology, Taiwan
- Manuel López Ibáñez, Complutense University of Madrid, Spain
- Aroua Taamallah, University of Sousse, Tunisia
- Anurag Deep, Indian Institute of Technology Bombay, India
- Chung-Yuan Hsu, National Pingtung University of Science and Technology, Taiwan
- Maha Khemaja, University of Sousse, Tunisia
- Gheorghita Ghinea, Brunel University London, England
- Liang-Yi Li, National Taiwan Normal University, Taiwan
- Ben Chang, National Central University, Taiwan
- Liz Bacon, Abertay University, Scotland
- Ismael Sagredo, International University of La Rioja, Spain
- Jianhua Wu, Central China Normal University, China
- Saurabh Mehta, Vidyalankar Institute of Technology, India
- Hiroyuki Mitsuhara, Tokushima University, Japan
- Alejandro Romero Hernández, Complutense University of Madrid, Spain
- Chang-Yen Liao, National Central University, Taiwan
- AhmedTlili, Beijing Normal University, China

C6 PC Members

- Wanwisa Wannapipat, Khon Kaen University, Thailand
- Hiroshi Miyashita, Tokyo Metropolitan Showa High School, Japan
- Elena Barcena, National Distance Education University, Spain
- Olga Viberg, KTH Royal Institute of Technology, Sweden
- Liliana Cuesta, University of La Sabana, Columbia
- Weichao Vera Chen, Baylor College of Medicine, United States
- Mark Pegrum, University of Western Australia, Australia
- Chia-Ling Hsieh, National Taiwan Normal University, Taiwan
- Daria Mizza Johns, Johns Hopkins University School of Advanced International Studies, United States
- Louise Ohashi, Meiji University, Japan
- Georgios Kormpas, Al Yamamah University, Saudi Arabi
- Maria Psychogiou, Athabasca University, Canada
- Valentina Morgana, Università Cattolica del Sacro Cuore, Milan
- Rustam Shadiev, Nanjing Normal University, China
- Debra Hoven, Athabasca University, Canada
- Chadia Mansour, Athabasca University, Canada
- Michał B. Paradowski, Institute of Applied Linguistics, University of Warsaw, Poland
- Carl Edlund Anderson, University of La Sabana, Columbia
- Antonie Alm, University of Otago, New Zealand

- Eric Hagley, Hosei University, Japan
- Phil Hubbard, Stanford University, United States
- Chien-Han Chen, Tamkang University, Taiwan
- Apostolos Koutropoulos, University of Masachusetts Boston, United States
- Sandra Gudiño-Paredes, Tecnológico De Monterrey, Mexico
- Yoshiko Goda, Kumamoto University, Japan
- Alex Boulton, University of Lorraine, France
- Di Zou, The Education University of Hong Kong, Hong Kong
- Takafumi Utashiro, Hokkai-Gakuen University, Japan
- Yuichi Ono, University of Tsukuba, Japan
- Misato Oi, Kyushu University, Japan
- Yasushige Ishikawa, Kyoto University of Foreign Studies, Japan
- Salomi Papadima-Sophocleous, Cyprus University of Technology, Cyprus
- Hayo Reinders, Unitec Institute of Technology, New Zealand
- Yanhui Han, The Open University of China, China
- Jiahang Li, Michigan State University, United States
- Adam Roarty, Rikkyo University, Japan
- Xin Chen, Indiana University, United States

C7 PC Members

- Mishra Shitanshu, Indian Institute of Technology Bombay, India
- Dan Kohen-Vacs, Holon Institute of Technology, Israel
- Yogendra Pal, NIIT University, India
- Gautam Biswas, Vanderbilt University, United States
- Marc Jansen, Hochschule Ruhr West, Germany
- Ivica Boticki, University of Zagreb, Croatia
- Chronis Kynigos, The National and Kapodistrian University of Athens, Greece
- Joke Voogt, University of Amsterdam, Netherlands
- Sun Daner, The Education University of Hong Kong, Hong Kong
- Jayakrishnan M. Warriem, Indian Institute of Technology, India
- Ahmed Mohammed, Leyte Normal University, Philippines
- Eran Gal, Holon Institute of Technology, Israel
- Jan Pawlowski, University of Applied Sciences Ruhr West, Germany
- Dan Klein, Holon Institute of Technology, Israel
- Tessy Cerratto-Pargman, Stockholm University, Sweden
- Andreas Lingnau, University of Applied Sciences Ruhr West, Germany
- Ulrich Hoppe, University of Duisburg-Essen, Germany
- Marcelo Milrad, Leyte Normal University, Philippines
- Ahmad Kamal, Linnaeus University, Sweden
- Ronen Hammer, Holon Institute of Technology, Israel
- Brendan Flanagan, Kyoto University, Japan
- Arriel Benis, Holon Institute of Technology, Israel
- Tamar Ronen Fuhrmann, Columbia University, United States
- Peter Seow, National Institute of Education, Singapore
- Winnie Wai Man Lam, The Education University of Hong Kong, Hong Kong
- Fredrik Hanell, Linnaeus University, Hong Kong
- Veenita Shah, Indian Institute of Technology Bombay, India
- Meital Amzalag, Holon Institute of Technology, Israel

- Chee-Kit Looi, National Institute of Education, Singapore
- Hayely Weigelt-Marom, Holon Institute of Technology, Israel
- Lee Shushing, National Institute of Education, Singapore

APSCE FELLOWS PROGRAM

Founded in 2019, the APSCE Fellowship recognizes outstanding members of the Asia-Pacific Society for Computers in Education (APSCE) in the field of computers in education. The title of APSCE fellow indicates, (1) Sustained and distinguished academic contributions to the advancement of research in the field of computers in education at the international level; (2) A strong track record in academic networking and services within the Asia-Pacific region.

The fellowship is for life, whose names shall be indicated on the APSCE website permanently. Furthermore, the APSCE fellows are entitled to complimentary lifetime voting APSCE memberships.

The number of new fellows named each year shall be capped at five (5). An APSCE Fellow must be an existing APSCE member in the year he or she is inducted.

The inaugural cohort of the APSCE Fellowship consists of the three existing APSCE Honorary Executive Committee (EC) members. Subsequently, the APSCE President, the APSCE Award Subcommittee Chair and the Honorary EC members formed the APSCE Fellow Committee to select additional fellows. After the first year (2019), the existing APSCE Fellows, the APSCE President and the Award Subcommittee Chair shall form the APSCE Fellow Committee each year to select new fellows. The APSCE President and the Award Subcommittee Chair are not eligible for APSCE Fellow inductions in the year in which they are serving as APSCE Fellow Committee members.

The full APSCE Fellowship guidelines is available on
https://apsce.net/download_data.php?filename=upfile/file/20201001/20201001020522_71700.pdf

The inaugural cohort of APSCE Fellows are (in alphabetical order):

- Tak-Wai CHAN (Taiwan)
- H. Ulrich HOPPE (Germany)
- Chee-Kit LOOI (Singapore)
- Riichiro MIZOGUCHI (Japan)

Two APSCE Fellows inducted in 2020 are (in alphabetical order):

- Gautam BISWAS (USA)
- Siu Cheung KONG (Hong Kong)

DISTINGUISHED RESEARCHER AWARD WINNER

Maria Mercedes (Didith) T. Rodrigo

Professor, Department of Information Systems and Computer Science, Ateneo Laboratory for the Learning Sciences (ALLS), Ateneo de Manila University,

Maria Mercedes (Didith) T. Rodrigo is a professor at the Department of Information Systems and Computer Science of the Ateneo de Manila University. The head of the Ateneo Laboratory for the Learning Sciences (ALLS), her research interests include artificial intelligence in education, technology in education, learning analytics, and affective computing. Under her leadership, ALLS has developed a number of mobile- and PC-based games for learning and augmented reality applications for informal learning in museums. Her team has also conducted eye tracking studies on novice programmer reading, tracing, and debugging skills.

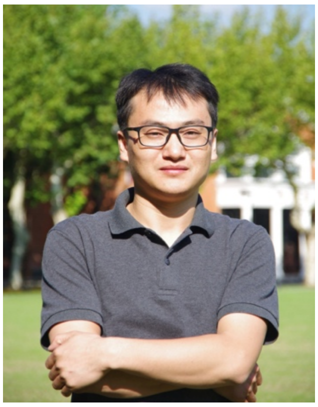
One of her most influential works is her contribution to the development of the Baker Rodrigo Ocumpaugh Monitoring protocol (BROMP), a quantitative field observation method that makes use of interval sampling to record student affect and behavior. Data from BROMP observations as labels to create student models. The BROMP protocol has been used in at least five countries and has become a de facto standard for classroom observations.

She continues to collaborate with colleagues from a variety of institutions including Dr. Ryan Baker and Dr. Jaclyn Ocumpaugh of the University of Pennsylvania's Center of Learning Analytics and Dr. H Chad Lane of the University of Illinois Urbana Champagne.

Dr. Rodrigo is an excellent mentor, as can be seen from a large number of postgraduate students she brings to the ICCE conferences. At ICCE 2019, the best overall paper award was given to one of the papers from her group.

Dr. Rodrigo has served on the Executive Committee of APSCE since 2014. She was the local organizing chair of ICCE 2018, International Program Coordinating Co-Chair of ICCE 2020, and International Program Coordinating Chair of ICCE 2021. She is also on the Executive Committee of the Artificial Intelligence in Education Society and is program chair of the International Artificial Intelligence in Education Conference 2022.

EARLY CAREER RESEARCHER AWARD WINNER (2021)



Bo Jiang

Associate Professor, Department of Educational Information Technology, East China Normal University, China

Dr. Bo Jiang is currently an associate professor in educational technology at the Department of Educational Information Technology, East China Normal University, China. He has a B.S. and M.S. in Computer Science. He received a Ph.D. degree in control science and engineering from Zhejiang University, China. His multidisciplinary background supports his research goal of creating new educational technologies to improve students' learning, and cultivating students' computational thinking as well as artificial intelligence literacy. His research interests include intelligent tutoring technologies, computational thinking education, artificial intelligence education. He has published over 40 publications (25 journal papers) in peer-reviewed journals, conferences, workshops, and book chapters. As project investigator, he received two grants from the National Natural Science Foundation of China. He is an Associate Editor of IEEE Transactions on Learning Technology, and International Journal of Bio-Inspired Computation. He is also elected to the Executive Committee of the Asia-Pacific Society on Computers in Education (APSCE) in 2018.

LAST TEN YEARS'

DISTINGUISHED RESEARCHER AWARD WINNERS

2020 - APSCE Distinguished Researcher Award
Wenli CHEN, Nanyang Technological University, Singapore

2015 - APSCE Distinguished Researcher Award
Lung-Hsiang WONG, Nanyang Technological University, Singapore

2014 - APSCE Distinguished Researcher Award
Hiroaki OGATA, Kyushu University, Japan

2011 - APSCE Distinguished Researcher Award
Chen-Chung LIU, National Central University, Taiwan
Antonija MITROVIC, University of Canterbury, New Zealand

LAST TEN YEARS'

EARLY CAREER RESEARCHER AWARD WINNERS

2020 - APSCE Early Career Researcher Award
Kaushal Kumar BHAGAT, Indian Institute of Technology, India

2019 - APSCE Early Career Researcher Award
Cheng-Jiu YIN, Kobe University, Japan

2018 - APSCE Early Career Researcher Award
Ting-Chia HSU, National Taiwan Normal University, Taiwan

2017 - APSCE Early Career Researcher Award
Jon MASON, Charles Darwin University, Australia

2015 - APSCE Early Career Researcher Award
Morris Siu-yung JONG, The Chinese University of Hong Kong, Hong Kong

Speakers of APSCE Webinar Series (December 2020 – November 2021)

Webinar 8: "Using Game Jam to Influence Cybersecurity Education"
Speaker: Lorie M. LIEBROCK & Amy KNOWLES, New Mexico Institute of Mining and Technology
Date: 17 December 2020, Thursday
Curated by: Special Interest Group on Educational Games and Gamification (EGG SIG), APSCE

Webinar 9: "Seamless Learning and Contextual Design"
Speaker: Christian GLAHN, Zurich University of Applied Sciences, Switzerland
Date: 24 February 2021, Wednesday
Curated by: Special Interest Group on Classroom, Mobile & Ubiquitous Technology Enhanced Learning (CUMTELL), APSCE

Webinar 10: " Panel on STEM Education in K-12: An International Perspective"
Speaker: Siu Cheung KONG (The Education University of Hong Kong, Hong Kong), Florence SULLIVAN (University of Massachusetts Amherst, USA), Morris Siu-Yung JONG (The Chinese University of Hong Kong, Hong Kong), Ting-Chia HSU (National Taiwan Normal University, Taiwan),& Timothy Ter Ming TAN (Nanyang Technological University, Singapore)
Date: 12 March 2021, Friday
Curated by: Special Interest Group on Computational Thinking Education & STEM Education (CTE&STEM), APSCE

Webinar 11: "Learning by Holding a Conversation with Computer Agents"
Speaker: Art GRAESSER, University of Memphis, USA
Date: 14 April 2021, Tuesday
Curated by: Special Interest Group on Educational Use of Problems/Questions in Technology-Enhanced Learning (EUPQ), APSCE

Webinar 12: "Beyond dashboards-Learning Analytics Architectures, Techniques and Applications"
Speaker: H. Ulrich HOPPE, University of Duisburg-Essen, Germany
Date: 29 April 2021, Thursday
Curated by: Special Interest Group on Learning Analytics and Educational Data Mining (LAEDM), APSCE

Webinar 13: "Speeches & Panel: How does new technology enhance vocabulary learning? Possibilities and issues"
Speaker: Yoshiko GODA (Kumamoto University, Japan), Yanjie SONG (The Education University of Hong Kong, Hong Kong), Nehal HASNINE (Hosei University, Japan), Vivian Wen-Chi WU (Asia University, Taiwan), Yuichi ONO (University of Tsukuba, Japan)
Date: 19 May 2021, Thursday
Curated by: Special Interest Group on Technology Enhanced Language Learning (TELL), APSCE

Webinar 14: "Understanding Learning Process in Classroom Interactions - Machine Learning for Discussion Forum Insights"
Speaker: Swapna GOTTIPATI, Singapore Management University, Singapore
Date: 8 June 2021, Tuesday
Curated by: Special Interest Group on Development of Information and Communication Technology in the Asia-Pacific Neighborhood (DICTAP), APSCE

Webinar 15: "Design and Practice of Gamified Online Course from The Knowledge Structure Embedding Perspective"
Speaker: Li WANG, Open University of China, China
Date: 20 August 2021, Friday
Curated by: Special Interest Group Educational Gamification and Game-based Learning (EGG), APSCE

Webinar 16: "Longitudinal Implications of Wheel Spinning and Productive Persistence"
Speaker: Ryan BAKER, University of Pennsylvania, USA
Date: 10 September 2021, Friday
Curated by: Special Interest Group on Artificial Intelligence in Education/Intelligent Tutoring Systems/Adaptive Learning (AIED/ITS/AL), APSCE

Webinar 17: "Design for Emergence: Conceptual and Technology Support for Student-Driven Knowledge Building"
Speaker: Tan DAO, Beijing Normal University, China; & Guangji YUAN, Nanyang Technological University, Singapore
Date: 27 September 2021, Thursday
Curated by: Special Interest Group on Computer-Supported Collaborative Learning & Learning Sciences (CSCL/LS), APSCE

Webinar 18: "Learning, Teaching, Teacher Professional Development and Policy in the Post-Pandemic World"
Speaker: Camille DICKSON-DEANE (University of Technology Sydney, Australia), Sharon RAVITCH (University of Pennsylvania, USA), Anvar SADAT (Kerala Infrastructure and Technology for Education, India), Jayakrishnan Madathil WARRIEM (Indian Institute of Technology Madras, India)
Date: 19 October 2021, Tuesday
Curated by: Special Interest Group on Practice-Driven Research, Teachers' Professional Development & Policies (PTP), APSCE

Webinar 19: "From Learning Object to Learning Cell, a Resource Organization Model for Ubiquitous Learning"
Speaker: Shengquan YU, Beijing Normal University, China
Date: 2 November 2021, Tuesday
Curated by: Special Interest Group on Advanced Learning Technologies (ALT), APSCE

KEYNOTE SPEAKERS

Kulthida TUAMSUK

Khon Kaen University, Thailand

Title:

Digital Learning Ecosystem for Transforming Classroom into Learning Community: Experiences from the Khon Kaen University Smart Learning Academy

Abstract: Keynote for the topic on digital learning ecosystem for transforming classroom into learning community will be presented based on the lesson learned from the Khon Kaen University (KKU) Smart Learning project which has been implementing at more than 200 junior high schools in the northeast of Thailand for 5 years. The presentation will cover main three topics: (1) initiation and background of KKU Smart Learning Academy; (2) KKU smart learning model, which is the principle and concept of learning competency development of students from the research and development of the research team in the project; and (3) overview of how the KKU smart learning model has been used at schools, and the making process of the digital learning ecosystem in classrooms that promote students' learning. Lastly, conclusion will be on the lessons we have learned from the work.

Keywords: Digital learning ecosystem, Smart learning, Junior high schools, Northeast Thailand, Learning community

Biography: Dr.Kulthida Tuamsuk is currently a Professor and Senior Researcher in Information Science at Khon Kaen University (KKU), Thailand. She is a founder and director of the KKU Smart Learning Academy which is one of the most well-known projects on the development of students' competencies at the secondary level in Thailand. This project is established based on the research works in multi-disciplinary done by researchers from various faculties in KKU. The KKU Smart Learning project has launched and implemented their innovative model (methods and products) for transforming the learning ecology in classroom, through the principled concepts: transforming the students' learning, by transforming the classroom and the teaching approach. The impacts of this project recognized at national level and accepted by the Ministry of Education.

Dr. Kulthida Tuamsuk is also specialized in Digital Humanities Research. Her research works in this field are well recognized and published in high quality international journals. She is a co-founder of a Consortium of iSchools in Asia Pacific (CiSAP), member of the Committee on Digital Humanities, iSchools Organization, and the former member of International Relation Affairs, the Association for Information Science and Technology (ASIS&T).

Pierre DILLENBOURG

Swiss Federal Institute of Technology, Switzerland
Title:
The classroom as a digital system

Abstract: Entering a modern car is like entering a computer with wheels, seats and windows. Similarly, entering a classroom is like entering a large digital system with chairs, windows and a board. The input devices of this system are not a keyboard and mouse, but an entire classroom equipped with sensors. The output device of this system is not a screen but a set of digital elements distributed in the class. The output is of course not a simple reflection of the input but input data are processed by multiple operators that aggregate, compare and visualize data. The resulting dashboards are used for monitoring the learners' progress in order to decide when and to whom to intervene. They are also used to compile data from the constructivist activities for supporting the debriefing phase, as well as to predict the completion time of an activity. Monitoring, debriefing and timing are central processes in classroom orchestration.

Biography: A former teacher in elementary school, Pierre Dillenbourg graduated in educational science (University of Mons, Belgium). He started his research on learning technologies in 1984. In 1986, he has been on of the first in the world to apply machine learning to develop a self-improving teaching system. He obtained a PhD in computer science from the University of Lancaster (UK), in the domain of artificial intelligence applications for education. He has been assistant professor at the University of Geneva. He joined EPFL in 2002. He has been the director of Center for Research and Support on Learning and its Technologies, then academic director of Center for Digital Education, which implements the MOOC strategy of EPFL (over 2 million registrations). He is full professor in learning technologies in the School of Computer & Communication Sciences, where he is the head of the CHILI Lab: "Computer-Human Interaction for Learning & Instruction ». He is the director of the leading house DUAL-T, which develops technologies for dual vocational education systems (carpenters, florists,…). With EPFL colleagues, he launched in 2017 the Swiss EdTech Collider, an incubator with 80 start-ups in learning technologies. He (co-)-founded 4 start-ups, does consulting missions in the corporate world and joined the board of several companies or institutions. In 2018, he co-founded LEARN, the EPFL Center of Learning Sciences that brings together the local initiatives in educational innovation. He is a fellow of the International Society for Learning Sciences. He currently is the Associate Vice-President for Education at EPFL.

Tiffany BARNES

NC State University, USA

Title:

Compassionate, Data-Driven Tutors for Problem Solving and Persistence

Abstract: Dr. Tiffany Barnes is a Distinguished Professor of Computer Science at North Carolina State University, and a Distinguished Member of the Association of Computing Machinery (ACM). Prof. Barnes is Founding Co-Director of the STARS Computing Corps, a Broadening Participation in Computing Alliance funded by the U.S.A. National Science Foundation. Her internationally recognized research program focuses on transforming education with AI-driven learning games and technologies, and research on equity and broadening participation. Her current research ranges from investigations of intelligent tutoring systems and teacher professional development to foundational work on educational data mining, computational models of interactive problem-solving, and design of computational thinking curricula. Her personalized learning technologies and broadening participation programs have impacted thousands of K-20 students throughout the United States.

Biography: Determining how, when, and whether to provide personalized support is a well-known challenge called the assistance dilemma. A core problem in solving the assistance dilemma is the need to discover when students are unproductive so that the tutor can intervene. This is particularly challenging for open-ended domains, even those that are well-structured with defined principles and goals. In this talk, I will present a set of data-driven methods to classify, predict, and prevent unproductive problem-solving steps in the well-structured open-ended domains of logic and programming. Our approaches leverage and extend my work on the Hint Factory, a set of methods that to build data-driven intelligent tutor supports using prior student solution attempts. In logic, we devised a HelpNeed classification model that uses prior student data to determine when students are likely to be unproductive and need help learning optimal problem-solving strategies. In a controlled study, we found that students receiving proactive assistance on logic when we predicted HelpNeed were less likely to avoid hints during training, and produced significantly shorter, more optimal posttest solutions in less time. In a similar vein, we have devised a new data-driven method that uses student trace logs to identify struggling moments during a programming assignment and determine the appropriate time for an intervention. We validated our algorithm's classification of struggling and progressing moments with experts rating whether they believe an intervention is needed for a sample of 20% of the dataset. The result shows that our automatic struggle detection method can accurately detect struggling students with less than 2 minutes

of work with 77% accuracy. We further evaluated a sample of 86 struggling moments, finding 6 reasons that human tutors gave for intervention from missing key components to needing confirmation and next steps. This research provides insight into the when and why for programming interventions. Finally, we explore the potential of what supports data-driven tutors can provide, from progress tracking to worked examples and encouraging messages, and their importance for compassionately promoting persistence in problem solving.

Gwo-Jen HWANG

National Taiwan University of Science and Technology, Taiwan

Title:

Applications and Research Issues of Artificial Intelligence in Education in the Mobile Era

Abstract: The advancement of artificial intelligence (AI) technologies has attracted the attention of researchers in the globe. However, it remains a challenging task for educational technology researchers to apply AI technologies to school settings, not to mention designing AIED (Artificial Intelligence in Education) studies. In this talk, Prof. Hwang is going to introduce the basic conceptions and applications of AI; following that, potential research issues of AIED in the mobile era are presented. In addition, several examples are given to demonstrate how AI can be used to promote teaching and learning outcomes. Finally, several approaches to designing and implementing AIED research are demonstrated.

Biography: Prof. Gwo-Jen Hwang is Chair Professor of Graduate Institute of Digital Learning and Education in National Taiwan University of Science and Technology. He serves as an editorial board member and a reviewer for more than 40 academic journals of educational technology and e-learning. Currently, he is the Editor-in-Chief /Lead Editor of Australasia Journal of Educational Technology (SSCI), International Journal of Mobile Learning and Organisation (Scopus, Q1), Journal of Computers in Education (Scopus, ESCI) and Computers & Education: Artificial Intelligence (Elsevier).

Prof. Hwang has published more than 700 academic papers, including more than 200 SSCI journal papers. Owing to his reputation in academic research and innovative inventions in e-learning, he received the annual most Outstanding Researcher Award from the National Science Council of Taiwan in the years 2007, 2010 and 2013. Moreover, in 2016, he was announced by Times Higher Education as being the most prolific and cited researcher in the world in the field of social sciences: https://www.timeshighereducation.com/news/ten-most-prolific-and-most-cited-researchers.

THEME-BASED INVITED SPEAKERS

Ana GIMENO

Universitat Politècnica de Valencia, Spain
Title:
Do Massive Open Online Language Courses (LMOOCs) Satisfy Learner Needs?

Abstract: Judging from what we hear and read, there seem to be as many supporters as detractors of Massive Open Online Courses (MOOCs). However, MOOCs are still a growing phenomenon and rely on technology to reach out to potential learners in populated cities as well as remote rural areas. Higher education in particular hasembraced this education "outlet" as a way to cater for an increasing demand for high quality online course materials to cover the needs of professionals who would like to engage in lifelong learning, and to satisfy the need to be at the forefront of educational developments and gain more international visibility. However, currently available MOOC platforms are in many respects limited in terms of courseware design and implementation as they are based on the template approach to software authoring. This limitation increases when we think of MOOCs that are intended for language learning – one of the most cognitively demanding disciplines learners can be confronted with. These MOOCs are commonly referred to as Language MOOCs or LMOOCs. Based on the Prof. Gimeno's experience in designing four upper-intermediate level MOOCs for learners of English as a Foreign Language, which have attracted over 200,000 learners to date from 258 different countries, she will discuss the findings deriving from over 17,000 learner responses to a survey conducted longitudinally over a period of two and a half years to shed light on some of the factors involved in learner motivation, expectations and learning styles. Additionally, as lack of guidance and scaffolding are factors that can lead to learner drop-outs, she will discuss the solutions that were implemented to overcome these deficiencies. In line with this, as some of the more challenging areas in LMOOC design relate to providing opportunities for learners to practise speaking and writing skills, she will discuss ways of designing activities to support learner interaction and communication, considering that these must satisfy learners who come from very different educational backgrounds and cultures.

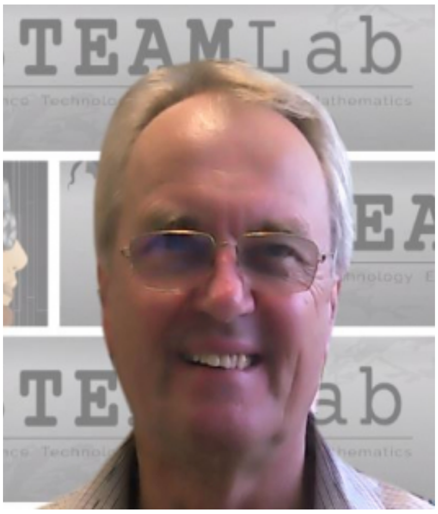
Biography: Ana Gimeno is Full Professor of English Language in the Department of Applied Linguistics at the Universitat Politècnica de València, Spain. She is Head of the CAMILLE Research Group, devoted to research in CALL and has been project manager of several funded multimedia CALL research and development projects that have led to the publication of a number of language courses in digital format. In 2016, she co-authored the first Spanish as a foreign language Massive Open Online Course (MOOC) delivered on the US-based edX platform, which has attracted over 350,000 learners from around the world and in 2018 she published the first upper-intermediate English edX MOOC, which has attracted over 200,000 learners. Ana Gimeno is Associate Editor of ReCALL (CUP) and serves on the Editorial Board of Computer-Assisted Language Learning Journal (Taylor and Francis), as well as being editor-in-chief of The EUROCALL Review. She wasPresident of the European Association for Computer–Assisted Language Learning (EUROCALL) for 6 years (2005-2011) and is currently President of the world organisation for computer-assisted language learning, WorldCALL (www.worldcall.org).

Baltasar FERNÁNDEZ-MANJÓN

Complutense University of Madrid, Spain
Title:
Systematizing Game Learning Analytics for Improving Serious Games Lifecycle

Abstract: Game learning analytics is the collection and analysis of user's gameplay interaction data to provide a better evidence-based insight on the educational process with serious games. The application of game learning analytics can provide a more data-driven scientific approach to improve all the steps of serious games' lifecycle. These steps include not only obtaining a better understanding about players learning and what actually happens when deploying a game in an educational scenario, but also enhancing the earlier steps of design, implementation and the overall quality of serious games. However, there is still a long way to go as learning analytics in games are not yet widespread and, in fact, there are very few serious games scientifically validated. The talk will introduce game learning analytics, their possible contributions to improve serious games lifecycle and the requirements (e.g. data standards, ethical considerations) for their systematization and generalization in real educational settings.

Biography: Dr.Baltasar Fernández-Manjón is a CS full professor (catedrático), and the leader of the e-learning research group e-UCM at the Complutense University of Madrid (UCM). Holder of the Telefonica-UCM Honorary Chair in Digital Education and Serious Games and IEEE Senior Member. Former Vice Dean of Research and Foreign Relationships at UCM. In 2010-11 he was Visiting Associate Professor at Harvard University and Visiting Scientist at LCS Massachusetts General Hospital. He has participated in a number of EU projects related with serious games technology and its application in different domains (e.g. H2020 RAGE and BEACONING, FP7 GALA, LLP SEGAN) where his group has been in charge of the learning analytics applied to the games (e.g. xAPI application profile for serious games, uAdventure for the creation of narrative and geolocation-based serious games). More info at https://www.e-ucm.es/people/balta/



Jon MASON

Charles Darwin University, Australia

Title:
Questioning and the Digital Environment

Abstract: To date, the digital environment has evolved rapidly around three key genres of innovation: search, social, and smart. If we take stock of where we're at right now then there's a mix of potential drivers of change – where technology can empower and enhance our experience and productivity … or it can frustrate and disrupt it. There is therefore both an upside and a downside in each innovation. When it comes to one of the most basic questions we all ask as children trying to make sense of things – why? – we don't yet have access to mature technologies that can scaffold this. This basic question is also fundamental to learning. This presentation will scan through some promising innovations that might inform the way we teach, learn, and research in the digital environment – and, not all these innovations are about technology. Moreover, 'questions that matter' are always contextual. Within these constraints, the question of how to facilitate questioning within the digital environment is the key theme explored in this presentation.

Biography: Dr.Jon Mason is an Associate Professor in Education in the College of Indigenous Futures, Education and Arts at Charles Darwin University (CDU), where he lectures in the broad area of digital technology in education. He also holds adjunct positions at Korea National Open University and East China Normal University. He first joined CDU in 2012 as Director of e-Learning for the Centre for School Leadership, Learning and Development and pursuing an earlier career at the nexus of government digital services, education, and international standardization. Since 2000 he has led delegations from Standards Australia to ISO/IEC JTC 1/SC36 and he has performed editorial roles for international projects, journals, and books. He is an elected member of the Executive Committee of the Asia Pacific Society for Computers in Education (APSCE) and serves on several journal editorial boards. His research encompasses most things where digital technology and learning intersect while also pursuing a keen interest in question formulation, sense-making and the role of wisdom in education.

PROGRAM AT A GLANCE

DSC: Doctoral Student Consortia **ECW**: Early Career Workshop
W: Workshop **WIPP**: Work-in-progress Posters

All times are in Bangkok Time Zone

November 22 (Monday)	November 23 (Tuesday)	November 24 (Wednesday)	November 25 (Thursday)	November 26 (Friday)
8:00 – 10:00 Tutorial / W02 / W03 / W04 / W06 / W09	8:00 – 10:00 DSC / W01 / W07 / W11 / W12	9:00 – 9:40 Opening Ceremony 9:40 – 10:00 DRA Speech	8:00 – 9:00 Keynote: Tiffany BARNES 9:00 – 9:40 Theme-based Speaker (Jon MASON) 9:40 – 10:00 Parallel Sessions 9:00 – 10:00 Parallel Sessions	9:00 – 10:00 Keynote: Gwo-Jen HWANG
10:00 – 10:20 Break		10:00 – 10:10 Break		
10:20 – 12:00 Tutorial / W02 / W03 / W04 / W06 / W09	10:20 – 12:00 DSC / W01 / W07 / W11 / W12	10:10 – 11:10 Keynote: Kulthida TUAMSUK 11:10 – 12:30 Parallel Sessions	11:00 – 12:30 Panel 3 10:10 – 11:20 Parallel Sessions 11:20 – 12:30 Parallel Sessions	10:10 – 11:10 Parallel Sessions 11:10 – 12:00 Closing Ceremony
12:00 – 1:00 Lunch		12:30 – 1:30 Lunch		12:00 – 1:30 Lunch
1:00 – 3:00 Student Wing Meeting W02 / W05 / W06 W08 / W09	1:00 – 3:00 ECW / W01 / W07 / W10 / W12	1:30 – 2:50 Panel 2 1:30 – 2:50 Parallel Sessions 2:50 – 3:00 Break	1:30 – 2:30 Keynote: Pierre DILLENBOURG 2:30 – 2:40 Break	1:00 – 3:30 EC Meeting
3:00 – 3:20 Break		3:00 – 3:40 Theme-based Speaker (Ana María GIMENO SANZ) 3:00 – 4:00 Parallel Sessions	2:40 – 3:40 Panel 1 2:40 – 4:00 Parallel Sessions	
3:20 – 5:00 SIG Leaders Meeting W02 / W05 / W06 W08 / W09	3:20 – 5:00 ECW / W01 / W07 / W10 / W12	4:00 – 5:30 Parallel Sessions	3:40 – 4:20 Theme-based Speaker (Baltasar FERNANDEZ-MANJON) 4:00 – 4:30 Parallel Sessions 4:30 – 5:30 Posters / WIPP	
	5:00 – 6:00 IPC Meeting			

November 22 (Monday)

Time	Room 1	Room 2	Room 3	Room 4	Room 5	Room 6
8:00 – 10:00	Tutorial - Kit-Build Concept Map: Effective Online Learning Through Concept Map Recomposition	Workshop 2	Workshop 3	Workshop 4	Workshop 6	Workshop 9
10:00 – 10:20	BREAK					
10:20 – 12:00	Tutorial - Kit-Build Concept Map: Effective Online Learning Through Concept Map Recomposition	Workshop 2	Workshop 3	Workshop 4	Workshop 6	Workshop 9
12:00 – 1:00	LUNCH					
1:00 – 3:00	Student Wing Meeting	Workshop 2	Workshop 8	Workshop 5	Workshop 6	Workshop 9
3:00 – 3:20	BREAK					
3:20 – 5:00	SIG Leaders Meeting	Workshop 2	Workshop 8	Workshop 5	Workshop 6	Workshop 9

November 23 (Tuesday)

Time	Room 1	Room 2	Room 3	Room 4	Room 5
8:00 – 10:00	DSC	Workshop 1	Workshop 7	Workshop 11	Workshop 12
10:00 – 10:20	BREAK				
10:20 – 12:00	DSC	Workshop 1	Workshop 7	Workshop 11	Workshop 12
12:00 – 1:00	LUNCH				
1:00 – 3:00	ECW	Workshop 1	Workshop 7	Workshop 10	Workshop 12
3:00 – 3:20	BREAK				
3:20 – 5:00	ECW	Workshop 1	Workshop 7	Workshop 10	Workshop 12
5:00 – 6:00	IPC Meeting				

November 24 (Wednesday)

Time	Room 1	Room 2	Room 3	Room 4
9:00 – 9:40	Opening Ceremony			
9:40 – 10:00	DRA Speech			
10:00 – 10:10	BREAK			
10:10 – 11:10	Keynote: Kulthida TUAMSUK			
11:10 – 12:30		EGG-1	ALT/LA/DI-1	PTP-1
12:30 – 1:30	LUNCH			
1:30 – 2:50	Panel 2 - Leveraging Student-Generated Ideas (SGI) to facilitate socio-constructivist learning and conceptual change	AIED/ITS-1	TELL-1	PTP-2
2:50 – 3:00	BREAK			
3:00 – 4:00	3:00 – 3:40 Theme-based Speaker 1: Ana María GIMENO SANZ (C6)	AIED/ITS-2	PTP-3	CUMTEL-1
4:00 – 5:30		CSCL/LS-1	ALT/LA/DI-2	CUMTEL-2

November 25 (Thursday)

Time	Room 1	Room 2	Room 3
8:00 – 9:00	Keynote: Tiffany BARNES		
9:00 – 10:00	9:00 – 9:40 Theme-based Speaker (Jon MASON)	AIED/ITS-3	CUMTEL-3
10:00 – 10:10	BREAK		
10:10 – 11:20	11:00 – 12:30 Panel 3 - Seeking quality in EdTech solutions: Perspectives from across the ecosystem	AIED/ITS-4	ALT/LA/DI-3
11:20 – 12:30		CSCL/LS-2	ALT/LA/DI-4
12:30 – 1:30	LUNCH		
1:30 – 2:30	Keynote: Pierre DILLENBOURG		
2:30 – 2:40	BREAK		
2:40 – 4:00	2:40 – 3:40 Panel 1 - The Role of Artificial Intelligence in STEM Education	EGG-2	TELL-2
4:00 – 4:30	3:40 – 4:20 Theme-based Speaker (Baltasar FERNANDEZ-MANJON)	PTP-4	ALT/LA/DI-5
4:30 – 5:30	Posters & WIPP		

November 26 (Friday)

Time	Room 1	Room 2	Room 3
9:00 – 10:00	Keynote: Gwo-Jen HWANG		
10:00 – 10:10	BREAK		
10:10 – 11:10		ALT/LA/DI-6	PTP-5
11:10 – 12:00	CLOSING CEREMONIES		
12:00 – 1:30	LUNCH		
1:00 – 3:30	EC Meeting		

CONFERENCE PROGRAM

DSC: Doctoral Student Consortia **ECW**: Early Career Workshop
W: Workshop **WIPP**: Work-in-progress Posters

F	Full Paper (20 mins. presentation, 5 mins. Q&A)
S	Short Paper (10 mins. presentation, 5 mins. Q&A)
ES	Extended Summary (10 mins presentation + 5 mins Q&A)
D	Demo Paper
BOPN	Best Overall Paper Nominee
BSPN	Best Student Paper Nominee
BTDPN	Best Technical Design Paper Nominee

All times are in Bangkok Time Zone

22 November 2021 (Monday)		
8:00 – 10:00	Tutorial: Kit-Build Concept Map: Effective Online Learning Through Concept Map Recomposition *Aryo PINANDITO, Hiroshima University, Japan, Universitas Brawijaya, Indonesia* *Didik Dwi PRASETYA., Hiroshima University, Japan, Universitas Negeri Malang, Indonesia* *Yusuke HAYASHI, Hiroshima University, Japan* *Tsukasa HIRASHIMA, Hiroshima University, Japan*	Room 1
	Workshop W02: The 14th Workshop on Technology Enhanced Learning by Posing/Solving Problems/Questions Workshop Organizers: *Yusuke HAYASHI, Hiroshima University, Japan* *Tsukasa HIRASHIMA, Hiroshima University, Japan* *Kazuaki KOJIMA, Teikyo University, Japan* *Tomoko KOJIRI, Kansai University, Japan* *Jon MASON, Charles Darwin University, Australia* *Antonija MITROVIC, University of Canterbury, New Zealand* *Fu-Yun YU. National Cheng Kung University, Taiwan* Complete list of papers: (*Note that the actual sequence of presentations will be determined by the workshop organizers*) #W02-01 F: Adopting Online Teaching and Learning Utilizing AI Technology Enhancements Throughout the COVID-19 Pandemic and Beyond Paul JENKINS, Nitin NAIK, Longzhi YANG #W02-02 D: A Learning Game on the Structure of Arithmetic Story by Chained Sentence Integration Kohei YAMAGUCHI, Yusuke HAYASHI, Tsukasa HIRASHIMA #W02-03 S: Exploring the Effects of the Collaborative and Cooperative Test-construction Strategies Chun-Ping WU #W02-04 S: Adopting No-Code Methods to Visualize Computational Thinking Derrick HYLTON, Shannon SUNG, Charles XIE #W02-05 F: Co-construction of Question-Led Inquiries Melvin FREESTONE, Jon MASON #W02-06 F: The Design and Effects of Online Contextual Student Generated Questions for English Grammar Learning Chih-Chung LIN, Fu-Yun YU	Room 2
	Workshop W03: Applications of Artificial Intelligence, Data Science and Intelligent Systems for Educational Research and Development (AIDS-ED) Workshop Organizers: *Assoc. Prof. Tossapon BOONGOEN, Mae Fah Luang University (MFU), Thailand* Co-organizers: *Prof. Qiang SHEN & Dr Changjing SHANG, Aberystwyth University, UK* *Asst.Prof. Shao-Chen CHANG, Yuan Ze University, Taiwan* *Prof. Taesu CHEONG, Korea University, South Korea* *Assoc.Prof. Kraisak KESORN Naresuan University, Thailand* *Asst.Prof. Chih-Hung CHEN, National Taichung University of Education, Taiwan*	Room 3

22 November 2021 (Monday)		
8:00 – 10:00	Workshop W03 Complete list of papers: (*Note that the actual sequence of presentations will be determined by the workshop organizers*) #W03-01 F: Adopting Online Teaching and Learning Utilizing AI Technology Enhancements Throughout the COVID-19 Pandemic and Beyond Paul JENKINS, Nitin NAIK, Longzhi YANG #W03-02 F: A Comparative Study of Missing Value Imputation Methods for Education Data Phimmarin KEERIN #W03-03 S: Learning Activities Diagnostic Model Based on Educational Data Mining of Online Social Media Behavior Khwunta KIRIMASTHONG, Pakphoom PROMMOO #W03-04 S: Improved Cluster Analysis for Graduation Prediction using Ensemble Approach Patcharaporn PANWONG, Natthakan IAM-ON, James MULLANEY	Room 3
	Workshop W04: Innovative Designs for Language Education : Transformative Technology Leadership Chair Organizer: *Dr. Bhornsawan INPIN, Mae Fah Luang University, Chiang Rai, Thailand* Co-Chair Organizers: *Xiong YUZHEN, Jinan University, China* *Mei-Rong Alice CHEN, National Taiwan University of Science and Technology, Taiwan* *Chi-Jen LIN, National Taiwan University of Science and Technology, Taipei, Taiwan* *Phirunkhana PHICHIENSATHIEN, Mae Fah Luang University, Chiang Rai, Thailand* Organizers: *Teeraparp PREDEEPORCH, Mae Fah Luang University, Chaing Rai, Thailand* *Nutdhavuth MEECHAIYO, Mae Fah Luang University, Chaing Rai, Thailand* Complete list of papers: (*Note that the actual sequence of presentations will be determined by the workshop organizers*) #W04-001: Developing and Evaluating a “Virtual Go Mode” Feature on an Augmented Reality App to Enhance Primary Students’ Vocabulary Learning Engagement Yanjie SONG, Yin YANG, Ka Man LUNG #W04-002: Thai-Chinese Interpretation Online Course Design-Identifying and accommodating learners’ needs Fang YUAN, Kanlaya KHAOWBANPAEW #W04-003: “How Do You Build a Literature Course?”: An Online Resource for Literature Curriculum Development Panida MONYANONT, Teeranuch ANURIT #W04-004: Design and Development of Video Instruction Utilizing a Flipped Classroom Model: Implementing Examples of Synonyms 不 and 没 Lalita RUKVICHAI, Natsarun LAKSANAPEETI #W04-005: The Development of Flipped Learning Model for Foreign Language Class Chun-Ye KIM #W04-007: A Challenge of Assistive Technology (AT) to the Needs of Visually Impaired (VIP) Learners in English Vocabulary Learning Phirunkhana PICHIENSATHIEN, Bhornsawan INPIN	Room 4
	Workshop W06: The 1st ICCE Workshop on EMBODIED Learning: Technology Design, Analytics & Practices Organizers: *Rwitajit MAJUMDAR, Kyoto University, Japan* *Aditi KOTHIYAL, Swiss Federal Institute of Technology Lausanne (EPFL) Switzerland.* *Prajakt PANDE, Roskilde University, Denmark.* *Shitanshu MISHRA, MGIEP UNESCO, India.* *Jayakrishnan Madathil WARRIEM, IIT Madras, India.* Complete list of papers: (*Note that the actual sequence of presentations will be determined by the workshop organizers*) #W06-01 F: Investigating Computer Designs for Grounded and Embodied Mathematical Learning Mitchell NATHAN, Candace WALKINGTON, Michael SWART #W06-02 S: Design for Remote Embodied Learning: The Hidden Village-Online Ariel FOGEL, Michael SWART, Jennifer SCIANNA, Matthew BERLAND, Mitchell NATHAN	Room 5

22 November 2021 (Monday)		
8:00 – 10:00	Workshop W06: The 1st ICCE Workshop on EMBODIED Learning: Technology Design, Analytics & Practices #W06-04 F: Embodied learning in makerspaces Ravi SINHA, Geetanjali DATE, Sanjay CHANDRASEKHARAN #W06-05 F: How Diseases Spread: Embodied Learning of Emergence with Cellulo Robots Hala KHODR, Jerome BRENDER, Aditi KOTHIYAL, Pierre DILLENBOURG #W06-06 S: An AI-enhanced Pattern Recognition Approach to Analyze Children's Embodied Interactions Ceren OCAK, Theodore J. KOPCHA, Raunak DEY #W06-07 S: Programming-RIO: Initiating Individuals into Computational Thinking using Real-world IoT Objects Spruha SATAVLEKAR, Shitanshu MISHRA, Ashutosh RAINA, Sridhar IYER #W06-09 S: Making mechanisms: how academic language mediates the formation of dynamic concepts Joseph SALVE, Pranshi UPADHYAY, Mashood K K, Sanjay CHANDRASEKHARAN #W06-10 F: Preparations for Multimodal Analytics of an Enactive Critical Thinking Episode Rwitajit MAJUMDAR, Duygu ŞAHIN, Yuanyuan YANG, Huiyong LI #W06-11 F: Teacher enaction: modeling how teachers build new mechanism concepts in students' minds Pranshi UPADHYAY, Joseph SALVE, Mashood K K, Sanjay CHANDRASEKHARAN	Room 5
	Workshop W09: The 10th International Workshop on ICT Trends in Emerging Economies (WICTTEE 2021) Chair: *Bo JIANG, East China Normal University, China* Co-chair: *Patcharin PANJABUREE, Mahidol University, Thailand* *Jayakrishnan M., Indian Institute of Technology Madras, India* *May TALANDRON-FELIPE, University of Science and Technology of Southern Philippines, Philippines* Consultant: *Su Luan WONG, Universiti Putra Malaysia, Malaysia* Complete list of papers: (*Note that the actual sequence of presentations will be determined by the workshop organizers*) #W09-01 S: Factors that Influence IT Students' Cyberchondria: Perspectives from the Philippines Don Erick BONUS, Ryan EBARDO #W09-02 S: The effectiveness of object-oriented-QR Monopoly in enhancing ice-breaking and education UX: A preliminary study Chien-Sing LEE, Kian-Wei LEE #W09-03 S: Mapping the Development of ICT from the Trend of Blended Learning: A Systematic Literature Review of the Blended Learning Trend in Education Lin WANGA, Muhd Khaizer OMARA, Noor Syamilah ZAKARIAA, & Nurul Nadwa ZULKIFLIB #W09-04 S: Students' Online Learning Experience during the COVID-19 pandemic: A Case Study at Universiti Putra Malaysia Su Luan WONGA, Mas Nida MD KHAMBARIB #W09-05 F: Transactional Distances During Emergency Remote Teaching Experiences Ma. Monica L. MORENO, Maria Mercedes T. RODRIGO, Johanna Marion R. TORRES, Timothy Jireh GASPAR, Jenilyn L. AGAPITO #W09-06 F: Exploring the differences in the cultivation of computational thinking in primary through meta-analysis based on the perspective of the contrast between the East and the West L Xiu GUAN, Guoxia WEIB, Bo JIANG, Xiang FENGA #W09-07 F: Designing Prototype Laryngo-App Using 3D Model in Anatomy of Larynx Poonyawee JIRARATTANAWAN, Ratchanon NOBNOP, Sujitra ARWATCHANANUKUL, Wimwipha SEEDET, Yootthapong TONGPAENG #W09-08 S: Comparing computational thinking in Scratch and non-Scratch Web design projects: A meta-analysis on framing and refactoring Chien-Sing LEE	Room 6

22 November 2021 (Monday)		
8:00 – 10:00	Workshop W09: The 10th International Workshop on ICT Trends in Emerging Economies (WICTTEE 2021) #W09-09 S: ICT used in Problem-Based Learning: Case study of a Thai University Nikorn RONGBUTSRI #W09-10 S: I Work to Learn: The Lived Experiences of Working Students in Online Learning During COVID-19 Ryan EBARDO, Santoso WIBOWO #W09-11 S: Developing a taxonomy of Edtech products for teachers: An integrated analysis from research literature and product landscape ISHIKAA, Gargi BANERJEE, Sahana MURTHY #W09-12 S: A Development of Instructional Video for Increasing Learners' Motivation and Content Mastery in Video Learning Environment Atima KAEWSA-ARD #W09-13 S: Design of Customizable Gamified Augmented Reality Apps: Towards Embracing Active Learning Mas Nida MD KHAMBARI, Dan WANG, Su Luan WONG, Priscilla MOSES, Mohd. Najwan MD. KHAMBARI, Rahmita Wirza O.K. RAHMAT, Fariza KHALID #W09-14 S: Promoting Transformative Citizenship in Diverse Society: An Appraisal of Massive Open Online Course as a Teaching Platform Anna Christi SUWARDI #W09-15 F: Perception of parents towards fun puzzle games in helping mild autistic children improve their computational thinking skills Chien-Sing LEE, Joey Nelson YATIM	Room 6
10:00 – 10:20	BREAK	
10:20 – 12:00	Continuation - Tutorial: Kit-Build Concept Map: Effective Online Learning Through Concept Map Recomposition (Room 1) Workshop W02 (Room 2) / Workshop W03 (Room 3) / Workshop W04 (Room 4) Workshop W06 (Room 5) / Workshop W09 (Room 6)	
12:00 – 1:00	LUNCH	
1:00 – 3:00	Student Wing Meeting: *Alwyn Vwen Yen LEE, Nanyang Technological University, Singapore*	Room 1
	Continuation - Workshop W02 (Room 2) / Workshop W06 (Room 5) / Workshop W09 (Room 6)	
	Workshop W08: Explorations in Online Teaching Modalities and Strategies Organizers: *Ma. Louise Antonette N. DE LAS PENAS, Ateneo de Manila University, Philippines* *Fr. Johnny C. GO, SJ, Ateneo de Manila University, Philippines* *Isabel Pefianco MARTIN, Ateneo de Manila University, Philippines* *Fr. Francis ALVAREZ, SJ, Ateneo de Manila University, Philippines* *Galvin Radley L. NGO, Ateneo de Manila University, Philippines* Complete list of papers: (*Note that the actual sequence of presentations will be determined by the workshop organizers*) #EOTMTS-001 F: Balancing the Pedagogical and Practical Concerns in Remote Higher Education: A Cyberethnography Jose Eos TRINIDAD, Samantha Joan ACKARY, Lyka Janelle PACLEB, Sophia Sue TABANAO, Jan Llenzl DAGOHOY #EOTMTS-004 F: Do Boredom, Escapism, Apathy, and Information Overload lead to Zoom Fatigue? Ryan EBARDO, Reynold PADAGAS, Hazel TRAPERO #EOTMTS-005 F: Exploring student behavior during Student-Generated Questions activities on Programming Learning Pham-Duc THO, Chih-Hung LAI, Thieu-Thi TAI #EOTMTS-006 F: Prelude to Full Online Learning: Educational Interventions from the Voice of the Customers Arlene Mae CELESTIAL-VALDERAMA, Albert A. VINLUAN, Joel B. MANGABA	Room 3

22 November 2021 (Monday)		
1:00 – 3:00	Workshop W05: The International Workshop on Learning Innovations in Science and Pre-Engineering Education (IWISPE) Chair: *Jintana WONGTA, King Mongkut's University of Technology Thonburi, Thailand* Co-Chair: *Feline Panas ESPIQUE, Saint Louis University, Philippines* *Cecilia A. MERCADO, Saint Louis University, Philippines* *Hideaki ABURATANI, National Institute of Technology (NIT) HQ, Japan* *Chiu-Lin LAI, National Taipei University of Education, Taiwan* *Ekapong HIRUNSIRISAWAT, King Mongkut's University of Technology Thonburi, Thailand* *Sukanlaya TANTIWISAWARUJI, King Mongkut's University of Technology Thonburi, Thailand* *Charoenchai WONGWATKIT, Mae Fah Luang University, Thailand* Consultant: *Kongkarn VACHIRAPANUNG, Learning Institute of King Mongkut's University of Technology Thonburi, Thailand* Complete list of papers: (*Note that the actual sequence of presentations will be determined by the workshop organizers*) #W05-001 S: Using Stellarium in Educating the Young Generation on Ancient Lanna Astronomy Pongsakorn PROMMING, Marut WONGTAPIN, Cherdsak SAELEE, Orapin RIYAPRAO #W05-002 S: Gaining Holistic Insight to Inthakin Festival via Stellarium Sky-mapping Veerapat SINTUPONG, Satanan SANPABOPIT, Suphakarn CHAISUK, Cherdsak SAELEE, Orapin RIYAPRAO #W05-003 F: Development an Online Workshop in Developing AI-Driven Mobile Application: Design and Analysis of Aidea Workshop 2021 Chirayu INTARATANOO, Jintana WONGTA, Natlada SIMASATHIEN, Panchika LORTARAPRASERT, Peerapat SUPASRI, Sattarin CHOOCHOUY #W05-004 S: The Design Process of STEM Learning Activities for Problem-Solving on the PM 2.5 Mask: The case of Primary School in Thailand Jirapipat THANYAPHONGPHAT, Wantana AREEPRAYOLKIJ, Suttikan LAKANUKAN #W05-005 S: ISOCHEM: Development of an Interactive 3D Game on the Web in Augmented Reality to Enhance Students' Learning of Isomers of Organic Chemistry Pannida PRASANSON, Jirapipat THANYAPHONGPHAT, Chatchadaporn PINTHONG #W05-006 S: A Development of Gamified Learning for Nursing Students' Public Health Investigation Process Pimpisa CHOMSRI, Pattranit SRISERM, Mullika MATRAKUL #W05-007 F: Targeting Chemistry Competencies on Plastic Circular Economy with Technology-assisted Citizen Inquiry: A Proposal of Learning Matrix Anggiyani Ratnaningtyas Eka NUGRAHENI, Banjong PRASONGSAP, Niwat SRISAWASDI #W05-008 F: Promoting Core Competencies of High School Biology through Citizen Inquiry Technology: A Case of Polluting Microplastics Arum ADITA, Chawadol SRIBOONPIMSUAY, Niwat SRISAWASDI	Room 4
3:00 – 3:20	BREAK	
3:20 – 5:00	SIG Leaders Meeting: *Weiqin CHEN, Oslo and Akershus University College of Applied Science, Norway*	Room 1
	Continuation - Workshop W02 (Room 2) / Workshop W08 (Room 3) / Workshop W05 (Room 4) Workshop W06 (Room 5) / Workshop W09 (Room 6)	

23 November 2021 (Tuesday)		
8:00 – 10:00	Doctoral Student Consortium Chair: *Morris JONG, The Chinese University Hong Kong, Hong Kong, China* Co-Chairs: *Hiroaki OGATA, Kyoto University, Japan* *Bo JIANG, East China Normal University, China* *Jayakrishnan Madathil WARRIEM, Indian Institute of Technology Madras, India*	Room 1

23 November 2021 (Tuesday)		
8:00 – 10:00	Workshop W01: The Applications of Information and Communication Technologies in Adult and Continuing Education Organizers: *Xibei XIONG, Guangxi Normal University, China* *Chunping ZHENG, Beijing University of Posts and Telecommunications, China* *Jyh-Chong LIANG, National Taiwan Normal University, Taiwan* *Min-Hsien LEE, National Taiwan Normal University, Taiwan* Complete list of papers: (*Note that the actual sequence of presentations will be determined by the workshop organizers*) #W01-001 F: Understanding Usage Continuance of Webinars among Professionals in the New Normal Ryan EBARDO, Jefferson COSTALES, Don Erick BONUS, Santoso WIBOWO, John Byron TUAZON, José Rizal University #W01-002 F: Research on Design Thinking and TPACK of Physical Education Pre-service Teachers Hungying LEE, Chingwei CHANG, Chiyang CHUNG #W01-003 F: A Review Study of the Application of Machine Translation in Education from 2011 to 2020 Yuyao ZHEN, Yaning WU, Guangming YU, Chunping ZHENG #W01-004 F: Online Interaction and Learning Engagement of Senior High School Students in a Less-Developed Region in China Jingyi WANG, Chunping ZHENG #W01-005 F: Perceived Teacher Support in Online Literature Reading: Scale Development, Validation, and Prediction of Continuous Reading Intention Yan SUN, Ying ZHOU, Jyh-Chong LIANG #W01-006 F: Exploring the online medical knowledge building in an university general education course Sheng-Han YANG, Jyh-Chong LIANG #W01-007 F: A Practical Study of Information Technology-Driven Teaching Reform of Innovation and Entrepreneurship in Higher Education Zhiming MENG, Xibei XIONG, Yu ZANG #W01-008 F: The Integration of Information Technology with Senior High English Reading Activities—A Case Study in Southwestern China Yanhua CHEN, Kehaoyu CHEN, Chenxi QIN	Room 2
	Workshop W07: The 5th Computer-Supported Personalized and Collaborative Learning Organizers: *Robin CHIU-PIN, Lin National Tsing Hua University, Taiwan* *Sherry Y. CHEN, National Central University, Taiwan and Brunel University, UK* *Gwo-Haur HWANG, National Yunlin University of Science and Technology, Taiwan* *Fu-Yun YU, National Cheng Kung University, Taiwan* *Wenli CHEN, National Institute of Education (NIE), Nanyang Technological University (NTU), Singapore* *Jitti NIRAMITRANON, Kasetsart University, Thailand* *Shu-Yuan TAO, Takming University of Science and Technology, Taiwan* *Hsiu-Ling CHEN, National Taiwan University of Science and Technology, Taiwan* Complete list of papers: (*Note that the actual sequence of presentations will be determined by the workshop organizers*) #W07-01 S: Design and Implementation of an iOS APP: Multimedia Interactive System and Items for Woodworking Teaching Chorng-Shiuh KOONG, Hung-Chang LIN, Chao-Chin WU, Chun-Hsien CHEN, Po-Huan LEE, Hsi-Chuan WANG #W07-02 F: The design of An Online Collaborative Orientation's Learning Activities to nurture Soft skills, Life skills and Self-directed learning Chanakan GROSSEAU, Jintana WONGTA, Arnon THONGSAW, Kongkarn VACHIRAPANANG #W07-03 F: Leveraging Context for Computer-Supported Student-generated Questions and EFL Learning in Grammar Instruction: Its Effects on Task Performance Wen-Wen CHENG, Fu-Yun YU	Room 3

23 November 2021 (Tuesday)		
8:00 – 10:00	Workshop W07: The 5th Computer-Supported Personalized and Collaborative Learning #W07-04 Poster: Mobile Learning System Combined with Adaptive Recommendation Mechanism--Taking outdoor learning activities of literature and history as an example Yu-Zhen DAI, Kai-Yi CHIN #W07-05 F: Academic Help-seeking Preference of Students during Online Flexible Learning May Marie P. TALANDRON-FELIPE, Gladys S. AYUNAR, Kent Levi A. BONIFACIO #W07-06 F: A Proposed Teacher Professional Development Program for Promoting Adult Teacher's TPACK in STEM Education Pawat CHAIPIDECH, Niwat SRISAWASDI #W07-07 Poster: The Impact of Inquiry-based Integrated STEM on Student's Perception of Learning Science and Computer Programming Chia-Jung CHANG, Shu-Yuan TAO	Room 3
	Workshop W11: The 5th International Workshop on Information and Communication Technology for Disaster and Safety Education (ICTDSE2021) Organizers: *Hisashi HATAKEYAMA, Tokyo Institute of Technology, Japan* *Hiroyuki MITSUHARA, Tokushima University, Japan* Advisory Member: *Ruggiero LOVREGLIO, Massey University, New Zealand* Complete list of papers: (*Note that the actual sequence of presentations will be determined by the workshop organizers*) #W11-002 S: Qualitative Evaluation of Information Display in a Regional Safety Map "Hamādo-map" Yasuhisa OKAZAKI, Tatsunari MEGURO, Hiroshi WAKUYA, Yukuo HAYASHIDA, Nobuo MISHIMA #W11-003 S: Towards Building Mental Health Resilience through Storytelling with a Chatbot Ethel ONG, Melody Joy GO, Rebecalyn LAO, Jaime PASTOR, Lenard Balwin TO #W11-004 F: Observing Evacuation Behaviors of Surprised Participants in Virtual Reality Earthquake Simulator Hiroyuki MITSUHARA, Itsuki TANIOKA, Masami SHISHIBORI #W11-005 S: A Proposal to Use Walk Rally Learning with Mystery Solving to Foster Attachment to Place and Understanding of Regional Characteristics Hisashi HATAKEYAMA, Masao MUROTA	Room 4
	Workshop W12: The 9th Workshop on Technology-Enhanced STEM Education (TeSTEM Workshop) Chair: *Charoenchai WONGWATKIT, Mae Fah Luang University, Thailand* Co-Chairs: *Niwat SRISAWASDI, Khon Kaen University, Thailand* *Patcharin PANJABUREE, Mahidol University, Thailand* *Ying-Tien WU, National Central University, Taiwan* *Sasithorn CHOOKAEW, King Mongkut's University of Technology North Bangkok, Thailand* Complete list of papers: (*Note that the actual sequence of presentations will be determined by the workshop organizers*) #W12-01 F: Trends of Engineering Design Process in STEM Education: A Systematic Review of the Evidence during 2017-2021 Teeratat SOPAKITIBOON, Surakit TUAMPOEMSAB, Sasithorn CHOOKAEW, Suppachai HOWIMANPORN #W12-02 S: Design and Development of an Augmented Reality Application for Learning a Serial-Link Robot Kinematics Sarut PANJANA, Tarinee TONGGOED #W12-03 S: Designing a Mobile Application to Promote Vocational Students' Learning in Basic Technical Drawings Course Tiptiya INTIP, Metha OUNGTHONG, Sasithorn CHOOKAEW #W12-04 S: Design and Development of a Low-Cost Robotic Platform for STEM Education of an Automatic Control System: Rotary-Type Double Inverted Pendulum Case Wiput TUVAYANOND, Piyanun RUANGURAI	Room 5

23 November 2021 (Tuesday)		
8:00 – 10:00	Workshop W12: The 9th Workshop on Technology-Enhanced STEM Education (TeSTEM Workshop) #W12-05 S: Developing a PLCs Experimental Kit through Role-playing for Students in Vocational Education Boonkert SONTIPAN, Suppachai HOWIMANPORN, Sasithorn CHOOKAEW #W12-07 S: Using Series Elastic Actuator as a Tool to Motivate Students Engineering Learning in STEM Education Piyanun RUANGURAI, Chaiyaporn SILAWATCHANANAI #W12-08 S: Designing a Project-based Plant to Enhance Mechatronics Students for Self-learning Embedded Control System Chaiyaporn SILAWATCHANANAI, Piyanun RUANGURAI #W12-09 S: Integrated Knowledge and Skills with Multi-Material Learning for Engineering Students during COVID-19 Suppachai HOWIMANPORN, Ornanong TANGTRONGPAIROS, Sasithorn CHOOKAEW #W12-10 S: Design Web-based Personalized Environment for Industrial Robots Learning Dechawut WANICHSAN, Konggrit PITANON, Sasithorn CHOOKAEW #W12-11 F: Development of a Gamified Number Line App for Teaching Estimation and Number Sense in Grades 1 to 7 Debbie Marie B. VERZOSA, Ma. Louise Antonette N. DE LAS PEÑAS, Maria Alva Q. ABERINB , Agnes D. GARCIANO, Jumela F. SARMIENTO, & Mark Anthony C. TOLENTINO #W12-12 F: For People and Planet: Evaluation of an Educational Mobile Game and Teacher Resource Pack Maria Mercedes T. RODRIGO, Johanna Marion R. TORRES, Janina Carla M. CASTRO, Abigail Marie T. FAVIS, Ingrid Yvonne HERRAS, Francesco U. AMANTE, Hakeem JIMENEZ, Juan Carlo F. MALLARI, Kevin Arnel MORA, Walfrido David A. DIY, Jaclyn Ting Ting VIDAL, Ma. Assunta C. CUYEGKENG #W12-13 F: Applying Outcomes-Based Learning in Mechatronics and Robotics Program: Case Study of Singburi Technical College Chatdanai SANEEWONGNAAYUTTAYA, Narongsak SANGNGOEN, Morakot KONGIN, Pattarapong PANHIRUN, Kittichai SAETUNG #W12-15 F: Case-based Professional Learning Course for Fostering Preservice Science Teachers' Technological Pedagogical and Content Knowledge of Inquiry with Mobile Game Phattaraporn PONDEE, Niwat SRISAWASDI	Room 5
10:00 – 10:20	BREAK	
10:20 – 12:00	Continuation - Doctoral Student Consortium (Room 1) Workshop W01 (Room 2) / Workshop W07 (Room 3) Workshop W11 (Room 4) / Workshop W12 (Room 5)	
12:00 – 1:00	LUNCH	
1:00 – 3:00	Early Career Workshop Chair: *Mas Nida MD. KHAMBARI, Universiti Putra Malaysia, Malaysia* Co-Chairs: *Ryan EBARDO, José Rizal University, Philippines* *Sharifah Intan Sharina SYED ABDULLAH, Universiti Putra Malaysia, Malaysia* *Patcharin PANJABUREE, Mahidol University, Thailand* Consultant: *Jon MASON, Charles Darwin University, Australia*	Room 1
	Continuation - Workshop W01 (Room 2) / Workshop W07 (Room 3) Workshop W12 (Room 5)	

23 November 2021 (Tuesday)		
1:00 – 3:00	Workshop W10: The 8th ICCE workshop on Learning Analytics and Evidence-based Education Organizers: *Huiyong LI, Kyoto University, Japan* *Rwitajit MAJUMDAR, Kyoto University, Japan* *Brendan FLANAGAN, Kyoto University, Japan* *Weiqin CHEN, Oslo Metropolitan University, Norway* *Hiroaki OGATA, Kyoto University, Japan* Complete list of papers: (*Note that the actual sequence of presentations will be determined by the workshop organizers*) #W10-001 S: Short Answer Questions Generation by Fine-Tuning BERT and GPT-2 Danny C.L. TSAI, Willy J.W. CHANG, Stephen J.H. YANG #W10-002 S: Exploring the Correlation between Students' Attention and Learning Performance Xin-Ping HUANG, Chung-Kai YU and Stephen J.H. YANG #W10-004 F: Identifying Students' Stuck Points Using Self-Explanations and Pen Stroke Data in a Mathematics Quiz Ryosuke NAKAMOTO, Brendan FLANAGAN, Kyosuke TAKAMI, Yiling DAI, Hiroaki OGATA #W10-005 S: Toward Educational Explainable Recommender System: Explanation Generation based on Bayesian Knowledge Tracing Parameters Kyosuke TAKAMI, Brendan FLANAGAN, Yiling DAI, Hiroaki OGATA #W10-006 S: Performance prediction and importance analysis using Transformer Akiyoshi SATAKE, Hironobu FUJIYOSHI, Takayoshi YAMASHITA, Tsubasa HIRAKAWA, Atsushi SHIMADA #W10-008 F: BEKT: Deep Knowledge Tracing with Bidirectional Encoder Representations from Transformers Zejie TIAN, Guangcong ZHENG, Brendan FLANAGAN, Jiazhi MI, Hiroaki OGATA #W10-010 F: Designing Nudges for Self-directed Learning in a Data-rich Environment Kinnari GATARE, Prajish PRASAD, Aditi KOTHIYAL, Pratiti SARKAR, Ashutosh RAINA, Rwitajit MAJUMDAR	Room 4
3:00 – 3:20	BREAK	
3:20 – 5:00	Continuation - Early Career Workshop (Room 1) Workshop W07 (Room 3) / Workshop W10 (Room 4) / Workshop W12 (Room 5)	
5:00 – 6:00	IPC Meeting	Room 1

24 November 2021 (Wednesday)		
9:00 – 9:40	OPENING CEREMONY	Room 1
9:40 – 10:00	DRA Speech: Maria Mercedes T. RODRIGO, Ateneo de Manila University, Philippines	Room 1
10:00 – 10:10	BREAK	
10:10 – 11:10	Keynote Speech 1: Kulthida Tuamsuk (C7) Moderator: *Thepchai SUPNITHI, National Electronics and Computer Technology Center, Thailand*	Room 1
11:10 – 12:30	EGG-1 Session Chair: *Maria Mercedes T. RODRIGO, Ateneo de Manila University* #6: Integrating Parsons Puzzles with Scratch - BOPN, BSPN (Full) Jeffrey BENDER, Bingpu ZHAO, Lalitha MADDURI, Alex DZIENA, Alex LIEBESKIND, Gail KAISER #3: A RECIPE for Teaching the Sustainable Development Goals (Short) Maria Mercedes T. RODRIGO, Walfrido David DIY, Abigail Marie FAVIS, Francesco AMANTE, Janina Carla CASTRO, Ingrid Yvonne HERRAS, Juan Carlo MALLARI, Kevin Arnel MORA, Johanna Marion R. TORRES and Ma. Assunta CUYEGKENG #20: Children Preference Analysis of A Mathematics Game- “Lily’s Closet" (Short) Wei Tung NIEN, Yi Chen WANG, Joni Tzuchen TANG #154: Tinkery: A Tinkerer’s Nursery for Problem Solving with Lego Mindstorms (Short) Ashutosh RAINA, Sridhar IYER, Sahana MURTHY	Room 2
11:10 – 12:30	ALT/LA/DI-1 Session Chair: *Tsukasa HIRASHIMA, Hiroshima University, Japan* #35: Developing a Generic Skill Assessment System Using Rubric and Checklists (Full) Makoto MIYAZAKI, Hiroyoshi WATANABE, Mieko MASAKA, Kumiko TAKAI #127: A Machine Learning Approach for Estimating Student Mastery by Predicting Feedback Request and Solving Time in Online Learning System (Full) Kannan N, Charles Y. C. YEH, Chih-Yueh CHOU, Tak-Wai CHAN #78: Classification of learning patterns and outliers using Moodle course material clickstreams and quiz scores (Short) Konomu DOBASHI, Curtis HO, Catherine FULFORD, Meng-Fen Grace LIN, Christina HIGA #86: Investigating the Tightness of Connection between Original Map and Additional Map in Extension Concept Mapping (Full) Didik PRASETYA, Aryo PINANDITO, Yusuke HAYASHI, Tsukasa HIRASHIMA	Room 3
	PTP-1 Session Chair: *Dan KOHEN-VACS, Holon Institute of Technology, Israel* #83: Mining Students’ Engagement Pattern in Summer Vacation Assignment (Full) Hiroyuki KUROMIYA, Rwitajit MAJUMDAR, Hiroaki OGATA #140: Alternative Approach for Evaluation Adapted for Times of Emergent Conditions (Full) Dan KOHEN-VACS, Meital AMZALAG #93: The Use of Video Conferencing Applications Facilitating Students' Behavioral Engagement during Synchronous Learning in the Time of Pandemic (Short) Mark Anthony ARIBON III #34: Facilitating collaborative learning among businesses, faculty, and students in a purely online setting (Short) Joseph Benjamin ILAGAN, Matthew Laurence UY, Vince Nathan KHO, Joselito OLPOC	Room 4

24 November 2021 (Wednesday)		
12:30 – 1:30	LUNCH	
1:30 – 2:50	Panel 2 (C2) - Leveraging Student-Generated Ideas (SGI) to facilitate socio-constructivist learning and conceptual change *Lung-Hsiang WONG, Chew Lee TEO, Longkai WU, Nanyang Technological University, Singapore*	Room 1
	AIED/ITS-1 Session Chair: *Ramkumar RAJENDRAN, Indian Institute of Technology Bombay, INDIA* #9: Investigating Engagement and Learning Differences between Native and EFL students in Active Video Watching (Full) Negar MOHAMMADHASSAN, Antonija MITROVIC #121: Authoring Tool for Semi-automatic Generation of Task-Oriented Dialogue Scenarios (Full) Emmanuel AYEDOUN, Yuki HAYASHI, Kazuhisa SETA #53: A System for Generating Student Progress Reports Based on Keywords (Short) Shumpei KOBASHI, Tunenori MINE #134: A coding mechanism for analysis of SRL processes in a technology enhanced learning environment (Short) Rumana PATHAN, Sahana MURTHY, Ramkumar RAJENDRAN	Room 2
	TELL-1 Session Chair: *May Marie P. TALANDRON-FELIPE, Ateneo De Manila University / University of Science and Technology of Southern Philippines, Philippines* #22: Comparison of English Comprehension among Students from different backgrounds using a Narrative-Centered Digital Game - BTDPN (Full) May Marie P. TALANDRON-FELIPE, Kent Levi A. BONIFACIO, Gladys S. AYUNAR, Maria Mercedes T. RODRIGO #88: Modelling the Relationship between English Language Learners' Academic Hardiness and Their Online Learning Engagement during the COVID-19 Pandemic (Short) Lin LUAN, Yanqing YI, Jinjin LIU #92: Design and Evaluation of a Game-based Language Learning Web Application for English Language Learners in Thailand (Short) Kornwipa POONPON, Wirapong CHANSANAM, Chawin SRISAWAT, Trinwattana POOCHANON #136: Proctored vs Unproctored Online Exams in Language Courses: A Comparative Study (Short) Mehmet ÇELIKBAĞ, Ömer DELIALIOĞLU	Room 3
	PTP-2 Session Chair: *Jon MASON, Charles Darwin University, Australia* #105: Low Adoption of Adaptive Learning Systems in Higher Education and How Can It Be Increased in Fully Online Courses (Full) Rhodora ABADIA, Sisi LIU #119: Co-designing for a healthy EdTech ecosystem: Lessons from the Tulna research-practice partnership in India (Full) Aastha PATEL, Chandan DASGUPTA, Sahana MURTHY, Rashi DHANANI #95: The M in STEM and Issues of Data Literacy (Short) Khalid KHAN, Jon MASON #146: Students with Disabilities and Digital Accessibility in Higher Education under COVID-19 (Short) Weiqin CHEN	Room 4

24 November 2021 (Wednesday)		
2:50 – 3:00	BREAK	
3:00 – 4:00	3:00 – 3:40 Theme-based Speaker 1: Ana María GIMENO SANZ, Universitat Politècnica de Valencia, Spain Moderator: *Yoshiko GODA, Kumamoto University, Japan*	Room 1
	AIED/ITS-2 Session Chair: *Yu LU, Beijing Normal University, China* #17: Improving knowledge tracing through embedding based on Metapath (Full) Chong JIANG, Wenbin GAN, Guiping SU, Yuan SUN, Yi SUN #29: Challenges to Applying Performance Factor Analysis to Existing Learning Systems (Short) Cristina MAIER, Ryan BAKER, Steve STALZER #30: Does Large Dataset Matter? An Evaluation on the Interpreting Method for Knowledge Tracing (Short) Yu LU, Deliang WANG, Penghe CHEN, Qinggang MENG	Room 2
	PTP-3 Session Chair: *Lucian NGEZE, Indian Institute of Technology Bombay, India* #102: Development and Preliminary Evaluation of an Online System in Support of a Student-Generated Testlets Learning Activity (Short) Fu-Yun YU #130: Learn to design (L2D): A TPD program to support teachers in adapting ICT learning materials to their local context through research-based strategies (Short) Gaurav JAISWAL, Sunita RASTE, Sahana MURTHY #132: Research on the Construction of Evaluation Indicators System of Pre-service teachers' Teaching Competency in Special Delivery Classroom (Short) Xiangchun HE, Peiliang MA, Xing ZHANG #152: From teaching to teacher training: Embedding important skills needed to develop a teacher trainer in cascaded Teacher Professional Development Programmes (Short) Lucian Vumilia NGEZE, Sridhar IYER	Room 3
	CUMTEL-1 Session Chair: *Brendan FLANAGAN, Kyoto University, Japan* #11: The Potential of Mobile Games in Improving Filipino and English Vocabulary among Children who are Non-native Speakers (Full) May Marie P. TALANDRON-FELIPE, Maria Mercedes T. RODRIGO #106: Analytics of Open-Book Exams with Interaction Traces in a Humanities Course - BOPN (Full) Rwitajit MAJUMDAR, Geetha BAKILAPADAVU, Jiayu LI, Hiroaki OGATA, Brendan FLANAGAN, Mei-Rong Alice CHEN	Room 4
4:00 – 5:30	CSCL/LS-1 Session Chair: *Fu-Yun YU, National Cheng Kung University, Taiwan* #42: Flip & Slack - Active Flipped Classroom Learning with Collaborative Slack Interactions (Full) Kyong Jin SHIM, Swapna GOTTIPATI, Yi Meng LAU #151: Designing Support for Productive Social Interaction and Knowledge Co-construction in Collaborative Annotation - BOPN, BSPN (Full) Xinran ZHU, Hong SHUI, Bodong CHEN #153: Theoretical and Practical Framework for a Multinational, Precollege, Peer Teaching Collaborative (Short) Eric HAMILTON, Danielle ESPINO, Seung LEE #5: A Measure to Cultivate Engaged Peer Assessors: A Validation Study on its Efficacy (Short) Yu-Hsin LIU, Kristine LIU, Fu-Yun YU	Room 2

24 November 2021 (Wednesday)		
4:00 – 5:30	ALT/LA/DI-2 Session Chair: *Hiroaki OGATA, Kyoto University, Japan* #48: Prior Knowledge on the Dynamics of Skill Acquisition Improves Deep Knowledge Tracing - BOPN (Full) Qiushi PAN, Taro TEZUKA #57: Blockchain in Education: Visualizations and Validating Relevance of Prior Learning Data (Short) Patrick OCHEJA, Brendan FLANAGAN, Rwitajit MAJUMDAR, Hiroaki OGATA #91: An AES System to assist teachers in grading Language Proficiency and Domain Accuracy using LSTM networks. (Short) Aditya SAHANI, Forum PATEL, Shivani MEHTA, Rekha RAMESH, Ramkumar RAJENDRAN #52: Analysis of the Answering Processes in Split-Paper Testing to Promote Instruction (Short) Shin UENO, Yuuki TERUI, Ryuichiro IMAMURA, Yasushi KUNO, Hironori EGI #ES-01: Improving Face-to-Face Communication Skills using Active Video Watching (Extended Summary) Ja'afaru MUSA, Antonija MITROVIC, Matthias GALSTER, Sanna MALINEN	Room 3
	CUMTEL-2 Session Chair: *Ryan EBARDO, Jose Rizal University, Philippines* #129: Human Factors in the Adoption of M-Learning by COVID-19 Frontline Learners (Full) Ryan EBARDO, Merlin Teodosia SUAREZ #137: Design Guidelines for Scaffolding Self-Regulation in Personalized Adaptive Learning (PAL) systems: A Systematic Review (Full) Vishwas BADHE, Gargi BANERJEE, Chandan DASGUPTA #120: EXAIT: A Symbiotic Explanation Education System (Short) Brendan FLANAGAN, Kyosuke TAKAMI, Kensuke TAKII, Dai YILING, Rwitajit MAJUMDAR, Hiroaki OGATA #114: Effects of virtual reality on students' creative thinking during a brainstorming session (Short) Mondheera PITUXCOOSUVARN, Victoria ABOU-KALIL, Hiroaki OGATA, Yohei MURAKAMI	Room 4

25 November 2021 (Thursday)		
8:00 – 9:00	Keynote Speech 4: Tiffany BARNES, North Carolina State University, United States Moderator: *Antonija MITROVIC, University of Canterbury, New Zealand*	Room 1
9:00 – 10:00	9:00 – 9:40 Theme-based Speaker: Jon MASON, Charles Darwin University, Australia Moderator: *Ramkumar RAJENDRAN, Indian Institute of Technology Bombay, India*	Room 1
	AIED/ITS-3 Session Chair: *Kazuhisa SETA, Osaka Prefecture University, Japan* #39: In-process Feedback by Detecting Deadlock based on EEG Data in Exercise of Learning by Problem-posing - BOPN, BTDPN (Full) Sho YAMAMOTO, Yuto TOBE, Yoshimasa TAWATSUJI, Tsukasa HIRASHIMA #68: Gaze- and Semantics-Aware Learning Material to Capture Learners' Comprehension Process (Short) Akio OKUTSU, Yuki HAYASHI, Kazuhisa SETA #68: Learning the Condition of Addition and Subtraction Word Problems by Problem-Posing based on Representation Conversion Model (Short) Yusuke HAYASHI, Natsumi TSUDAKA, Kengo IWAI, Tsukasa HIRASHIMA	Room 2

25 November 2021 (Thursday)		
9:00 – 10:00	CUMTEL-3 Session Chair: *Aungtinee KITTIRAVECHOTE, Bansomdejchaopraya Rajabhat University, Thailand* #55: Comparison of Experts and Novices in Determining the Gravitational Acceleration using Mobile Phone with Phyphox Application (Short) Aungtinee KITTIRAVECHOTE, Thanida SUJARITTHAM #116: Karyotype: An interactive learning environment for reasoning and sense making in genetics through a case-based approach (Short) Sunita RASTE, Anurag DEEP, Sahana MURTHY #117: Design and Deployment of a Mobile Learning Cloud Network to Facilitate Open Educational Resources for Asynchronous Learning (Short) Joselito Christian Paulus VILLANUEVA, Mark Anthony MELENDRES, Catherine Genevieve LAGUNZAD, Nathaniel Joseph LIBATIQUE #147: TinkerBot: A Semi-Automated Scaffolding Agent as a Companion for Tinkering. (Short) Shruti JAIN, Ashutosh RAINA, Sridhar IYER	Room 3
10:00 – 10:10	BREAK	
10:10 – 11:20	AIED/ITS-4 Session Chair: *Ethel ONG, De La Salle University, Philippines* #65: An Improved Model to Predict Student Performance Using Teacher Observation Reports - BSPN Menna FATEEN, Kyouhei UENO, Tsunenori MINE #23: Diverse Linguistic Features for Assessing Reading Difficulty of Educational Filipino Texts (Short) Joseph Marvin IMPERIAL, Ethel ONG #41: Using Qualitative Data from Targeted Interviews to Inform Rapid AIED Development (Short) Jaclyn OCUMPAUGH, Stephen HUTT, Juliana Ma. Alexandra L. ANDRES, Ryan BAKER, Gautam BISWAS, Nigel BOSCH, Luc PAQUETTE, Anabil MUNSHI #128: Reflection Support Environment for Creative Discussion Based on Document Semantics and Multimodal Information (Short) Atsuya SHONO, Yuki HAYASHI, Kazuhisa SETA	Room 2
	ALT/LA/DI-3 Session Chair: *May Kristine Jonson CARLON, Tokyo Institute of Technology, Japan* #64: A Thematic Summarization Dashboard for Navigating Student Reflections at Scale (Full) Yuya ASANO, Sreecharan SANKARANARAYANAN, Majd SAKR, Christopher BOGART #43: Learning Analytics Dashboard Prototype for Implicit Feedback from Metacognitive Prompt Responses (Short) May Kristine Jonson CARLON, Jeffrey CROSS #62: Analysing reachable and unreachable codes in App Inventor programs for supporting the assessment of computational thinking concepts (Short) Siu Cheung KONG, Chun Wing POON, Bowen LIU #96: Conceptual Level Comprehension Support of The Object-Oriented Programming Source-Code Using Kit-Build Concept Map (Short) Nawras KHUDHUR, Pedro Gabriel Fontales FURTADO, Aryo PINANDITO, Shimpei MATSUMOTO, Yusuke HAYASHI, Tsukasa HIRASHIMA	Room 3
11:00 – 12:30	Panel 3 - Seeking quality in EdTech solutions: Perspectives from across the ecosystem Sahana MURTHY, Indian Institute of Technology Bombay, India	Room 1
11:20 – 12:30	CSCL/LS-2 Session Chair: *Priscilla MOSES, Universiti Tunku Abdul Rahman, Malaysia* #113: Laboratory Study on ICAP Interventions for Interactive activity: Investigation Based on Learning Performance (Full) Shigen SHIMOJO, Yugo HAYASHI #38:STEM and Non-STEM Students' Perception towards Work Environment and Career Prospect (Short) Priscilla MOSES, Tiny Chiu YUEN TEY, Phaik Kin CHEAH	Room 2

25 November 2021 (Thursday)		
11:20 – 12:30	CSCL/LS-2 #79:Fostering conceptual change in software design (Short) Lakshmi T G, Sridhar IYER #87:The Effectiveness of Collaborative Concept Map Recomposition and Discussion with Kit-Build Concept Map in Online Learning (Short) Aryo PINANDITO, Didik PRASETYA, Nawras KHUDHUR, Yusuke HAYASHI, Tsukasa HIRASHIMA	Room 2
	ALT/LA/DI-4 Session Chair: *Paraskevi TOPALI, GSIC-EMIC Research Group, Spain* #144: Profiling Student Learning from Q&A Interactions in Online Discussion Forums - BSPN (Full) De Lin ONG, Kyong Jin SHIM, Swapna GOTTIPATI #82: Supporting MOOC Instructors in the Identification of Learner Problems Framed within the Learning Design (Short) Paraskevi TOPALI, Alejandro ORTEGA-ARRANZ, Alejandra MARTÍNEZ-MONÉS, Sara VILLAGRÁ-SOBRINO, Juan Ignacio ASENSIO-PÉREZ, Yannis DIMITRIADIS #124: Automatic classification of MOOC forum messages to measure the quality of peer interaction (Short) Urvi SHAH, Richa RAMBHIA, Prakruti KOTHARI, Rekha RAMESH, Gargi BANERJEE #150: Identifying and Comparing Topic Categories and Interaction Features in MOOC Discussions Supported by Danmaku (Short) Bo YANG	Room 3
12:30 – 1:30	LUNCH	
1:30 – 2:30	Keynote Speech 3: Pierre DILLENBOURG, Swiss Federal Institute of Technology, Switzerland Moderator: *Kate THOMPSON, Queensland University of Technology, Australia*	Room 1
2:30 – 2:40	BREAK	
2:40 – 3:40	Panel 1 - The Role of Artificial Intelligence in STEM Education *Siu Cheung KONG, The Education University of Hong Kong, Hong Kong*	Room 1
2:40 – 4:00	EGG-2 Session Chair: *Jonathan DL CASANO, Ateneo de Manila University, Philippines* #50: Tactical Knowledge Acquisition Support System from Play Videos of Esports Experts (Full) Minato SHIKATA, Tomoko KOJIRI #103: Robot with Embodied Interactive Modes as a Companion Actor in Journey of Digital Situational Learning Environment and its Effect on Students' Learning Performance - BTDPN (Full) Vando Gusti AL HAKIM, Su-Hang YANG, Jen-Hang WANG, Chiu-Chen YEN, Lung YEH, Gwo-Dong CHEN #8: Designing Games for Stealth Health & Healthy Lifestyle Education (Short) Nilufar BAGHAEI #107: Xiphias: Using A Multidimensional Approach Towards Creating Meaningful Gamification-Based Badge Mechanics (Short) Jonathan CASANO, Jenilyn AGAPITO, Nicole Ann TOLOSA	Room 2
	TELL-2 *Session Chair: Jingjing LIAN, Peking University, China* #36: A Quasi-experimental Study of University English Learners' Engagement in a Flipped Classroom - BOPN, BSPN (Full) Jingjing LIAN, Jiyou JIA #26: Integrating E-learning into Self-regulated Learning Instruction: A Holistic Flipped Classroom Design of a Classical Chinese Reading Intervention Program (Short) Kit Ling LAU #84: A Flipped Model of Active Reading Using Learning Analytics-enhanced E-book Platform (Short) Yuko TOYOKAWA, Rwitajit MAJUMDAR, Louis LECAILLIEZ, Hiroaki OGATA #100: The Learning Potential of Online Student-Generated Questions Based on Given Graphics for English Language Learning (Short) Fu-Yun YU	Room 3

25 November 2021 (Thursday)		
3:40 – 4:20	Theme-based Speaker: Baltasar FERNANDEZ-MANJON, Complutense University of Madrid (UCM) Spain Moderator: *Rita KUO, New Mexico Institute of Mining and Technology, USA*	Room 1
4:00 – 4:30	PTP-4 *Session Chair: Ma. Monica MORENO, Ateneo de Manila University, Philippines* #19: Cura Personalis: Institutionalizing Compassion During Emergency Remote Teaching (Short) Ma. Monica MORENO, Ma. Mercedes RODRIGO, Johanna Marion TORRES, Timothy Jireh GASPAR, Jenilyn AGAPITO #81: Educational leadership and children's resilience: German and Polish schools during COVID-19 (Short) Paulina BURKOT, Amy SEPIOŁ, Nataliia DEMESHKANT	Room 2
	ALT/LA/DI-5 *Session Chair: Yu LU, Beijing Normal University, China* #7: SLP: A Multi-Dimensional and Consecutive Dataset from K-12 Education(Short) Yu LU, Yang PIAN, Ziding SHEN, Penghe CHEN, Shengquan YU #143: Evaluation of a Motion Capture and Virtual Reality Classroom for Secondary School Teacher Training. (Short) Sandra ALONSO, Daniel LÓPEZ, Andrés PUENTE, Alejandro ROMERO, Ibis ALVAREZ, Borja MANERO	Room 3
4:30 – 5:30	Posters / Work-in-progress Posters	Room 2

26 November 2021 (Friday)		
9:00 – 10:00	Keynote Speech 2: Gwo-Jen HWANG, National Taiwan University of Science and Technology, Taiwan Moderator: *Jingyun WANG, Durham University, UK*	Room 1
10:00 – 10:10	BREAK	
10:10 – 11:10	ALT/LA/DI-6 *Session Chair: Nguyen-Thinh LE, Humboldt-Universität zu Berlin, Germany* #54: From Hello to Bye-Bye: Churn prediction in English Language Learning App (Full) Daevesh SINGH, Rumana PATHAN, Gargi BANERJEE, Ramkumar RAJENDRAN #126: How Can Pedagogical Agents Detect Learner's Stress? (Full) Nguyen-Thinh LE, Melanie BLECK, Niels PINKWART	Room 2
	PTP-5 *Session Chair: Siu Cheung KONG, The Education University of Hong Kong, Hong Kong* #60: From Mathematical Thinking to Computational Thinking: Use Scratch Programming to Teach Concepts of Prime and Composite Numbers (Full) Siu Cheung KONG, Wai Ying KWOK #112: GUI Based System for Effortless Program Visualization Creation Using Time Series Information - BOPN, BTDPN (Full) Koichi YAMASHITA, Miyu SUZUKI, Satoru KOGURE, Yasuhiro NOGUCHI, Raiya YAMAMOTO, Tatsuhiro KONISHI, Yukihiro ITOH	Room 3
11:10 – 12:00	CLOSING CEREMONY	
12:00 – 1:30	LUNCH	
1:00 – 3:30	EC Meeting	

POSTERS (P) & WORK-IN-PROGRESS POSTERS (WIPP)

25 November 2021 (Thursday) 4:30 – 5:30 PM, Room 2

P: Poster
WIPP: Work-in-progress Posters

#40 P: Viewpoint Transformation Training System Based on Discovery of Relationships Between Objects
Kota KUNORI and Tomoko KOJIRI

#47 P: Support System for Understanding Intention in Communication Using Diagrams
Koushi UEDA and Tomoko KOJIRI

#49 P: Chinese Grammatical Error Detection Using Adversarial ELECTRA Transformers
Lung-Hao LEE, Man-Chen HUNG, Chao-Yi CHEN, Rou-An CHEN, Yuen-Hsien TSENG

#69 P: Presentation Scenario Design Support System That Prompts Awareness of Other Viewpoints
Kazumi MASAKADO, Yuki HAYASHI, Kazuhisa SETA

#71 P: Visualization of Topics and Logical Development Based on Reader's Understanding of Inter-sentence Relations for Reading Support
Yuki OKANIWA, Tomoko KOJIRI

#28 P: Explore the contribution of learning style for predicting learning achievement and its relationship with reading learning behaviors
Fuzheng ZHAO, Bo JIANG, Juan ZHOU, Chengjiu YIN

#16 P: Research on the application of College Students' online learning cognitive engagement evaluation
Yonghong WANG, Xiangchun HE

#118 P: Integration of Programming-based Tasks into Mathematical Problem-based Learning
Zhihao CUI, Oi-Lam NG, Morris JONG

#149 P: Web-Based Engineering Design Activity in Biology: An Assessment on the Demonstration of Higher-Order Thinking Skills
Ma. Andrea Claire CARVAJAL

#32 P: A Mixed Study to Understand Taiwanese Children's Preference for A Mobile Game
Yi Chen WANG, Wei Tung NIEN, Joni Tzuchen TANG

#133 P: The impact of Augmented Reality applied in Vocabulary Learning and Use on Elementary EFL School Students
Chin-Huang LIAO, Wen-Chi Vivian WU, Tin-Chang CHANG, Chang-Hung LEE

#13 P: Birds of Paradise: A Game on Urban Bird Biodiversity Conservation
Jamielyn Mae VILLANUEVA, Patricia Vianne LEE, James Matthew CUARTERO

#46 P: Suggestions for special education teachers to practice spherical image-based virtual reality instruction in classrooms: a case study
Kun-Hung CHENG

#51 P: Computational Fluency and the Digital Divide in Japanese Higher Education
Luc GOUGEON, Jeffrey CROSS

#70 P: Improvement of Teaching Based on the E-book Reader Logs: A Case Study at High School Math Class in Japan
Taro NAKANISHI, Hiroyuki KUROMIYA, Rwitajit MAJUMDAR, Hiroaki OGATA

#21 P: Composition class using a system that encourages self-review —Focus on second language learning—
Yan ZHAO, Haruhiko TAKASE, Hidehiko KITA

#89 P: Examining the effects of automatic speech recognition technology on learners' lexical diversity
Michael JIANG, Morris JONG, Wilfred LAU

#139 P: Technology Integration in a Communicative English Classroom
Ruth Z. HAUZEL

#WIPP-01: A Trial Study on Restraint of Mind-Wandering while Viewing Educational Videos by Adjusting Biofeedback Operations
Toru NAGAHAMA, Naoki NOSE, Issaku KAWASHIMA, Yusuke MOTITA

#WIPP-02: Narrative Discourse Structure Creation Support System for Reflecting Theme and Emotional Impression
Atsushi ASHIDA, Masataka TOKUMARU, Tomoko KOJIRI

#WIPP-03: Visualization Method of Movement of Teachers and Students in Classroom using OpenPose
Misato FUTATSUISHI, Izumi HORIKOSHI, Yasuhisa TAMURA

#WIPP-04: Development of Mapping Function between Variable Value and Object Properties for Program Behavior Visualization Tool TEDViT
Hiroki SOMA, Satoru KOGURE, Yasuhiro NOGUCHI, Koichi YAMASHITA, Raiya YAMAMOTO, Tatsuhiro KONISHI, Yukihiro ITOH

#WIPP-05: Extraction Method of Characteristics of Important Body Shapes and their Training Order for Motor Skill Acquisition
Jinya SUMIZAKI, Tomoko KOJIRI

#WIPP-06: Instruction Support System Using Impasse Detector and Major Failure Diagnoser for Programming Exercises
Tomoki IKEGAME, Yasuhiro NOGUCHI, Satoru KOGURE, Koichi YAMASHITA, Raiya YAMAMOTO, Tatsuhiro KONISHI, Yukihiro ITOH

#WIPP-07: Augmented Reality (AR) Analytics to Investigate Motor Skills for Crossing the Midline
Manjeet SINGH, Shaun BANGAY, Atul SAJJANHAR

#WIPP-08: Motivating Ethnic Minority Students in Hong Kong to Learn Chinese Culture with EduVenture VR
Morris Siu-Yung JONG, Nelson NG, Eric LUK, Jessie LEUNG, Michael Yi-Chao JIANG, Darwin LAU, Chin-Chung TSAI

#WIPP-09: An Augmented Reality Experience for Generating New Audiences for Spanish Dance
Alejandro ROMERO-HERNANDEZA, Lara MARIN, Borja MANERO

#WIPP-10: The Development and Evaluation of an Online Educational Game Integrated with Real Person-NPC mechanism for History Learning
Shu-Wei LIU, Hung-Yu CHAN, Huei-Tse HOU

#WIPP-11: Role of Preparation using Mobile Application for Summary-speaking Task in Face-to-face English-Speaking Pair Work
Kae NAKAYA, Masao MUROTA

#WIPP-12: Proficiency, learning strategies, and logging behaviors on the dictation training courseware
Yuichi Ono

TABLE OF CONTENTS

Investigating Engagement and Learning Differences between Native and EFL students in Active Video Watching

Negar MOHAMMADHASSAN* **& Antonija MITROVIC**
University of Canterbury, New Zealand
*negar.mohammadhassan@pg.canterbury.ac.nz

Abstract: Video-based learning (VBL) requires good listening and reading comprehension skills, which could be challenging for English as a foreign language (EFL) students. In this paper, we investigate the differences between EFL and Native English speakers in a VBL platform called AVW-Space, in order to identify potential interventions that would be helpful for EFL students. AVW-Space provides note-taking, peer-reviewing, visualisations and personalised nudges to support engagement in VBL. Although previous studies on AVW-Space showed these supports were effective for increasing engagement, we discovered significant differences in learning outcomes and engagement between EFL/Native students, which stem from different learning strategies, background knowledge and language barriers. This research contributes to using learning analytics to understand better the differences between EFL and Native students, and providing more specialised support for EFL students in VBL.

Keywords: Video-based learning, English as a foreign language students, Learning analytics, Personalised support, Student equity and inclusion

1. Introduction

Using videos for learning has become very popular due to platforms such as YouTube and Massive Open Online Courses (MOOCs). Online video-based learning (VBL) platforms provide easy access to educational materials to people from all over the world. VBL is an effective method of learning since it combines visual, textual and auditory modes. However, VBL requires good listening and reading comprehension skills, which could make VBL challenging for English as a foreign language (EFL) students. Thus, in order to have adaptive VBL platforms (Giannakos, Sampson, & Kidziński, 2016), the needs and abilities of EFL students should be taken into account. To identify what kind of support would be helpful for EFL students in VBL, we first need to understand the differences between EFL and Native students in terms of their learning strategies, engagement and learning outcomes.

This research investigates the differences between EFL/Native students in AVW-Space. AVW-Space (Mitrovic et al., 2016) is an online VBL platform that supports engagement via note-taking, micro-scaffolds for reflection, reviewing comments made by peers, visual learning analytics and personalised nudges (Mitrovic et al., 2017; Mitrovic et al., 2019). In AVW-Space, the student can watch a video, pause it to write a comment and tag the comment with one of the aspects defined by the teacher, which encourage reflection. An early study with AVW-Space (Mitrovic et al., 2017) in the context of presentation skills found that students who commented on videos learned more than their peers who watched videos passively. In order to support active video watching, personalised Reminder nudges were added to AVW-Space to encourage students to write comments and use various aspects (Mitrovic et al., 2019). Later on, personalised Quality nudges were introduced in order to encourage students to write high-quality comments. The Quality nudges automatically assess the quality of comments students write (Mohammadhassan et al., 2020) and guide students toward critical thinking and self-reflection (Mohammadhassan et al., 2021). AVW-Space also supports social learning: the student can review and rate comments made by their classmates using the rating options that the teacher defines. However, AVW-Space has no specialised support for EFL students. Thus, this research investigates the differences between EFL and Native students who used AVW-Space for learning presentation skills. Based on the findings, we discuss potential modifications for EFL students. The paper addresses the following research questions:

RQ1: Do Native and EFL students have different self-reported learning strategies?
RQ2: Is there a difference in engagement and learning outcomes for EFL/Native students?
RQ3: What are the differences in the subjective opinions of EFL/Native students on interactions with AVW-Space?

2. Related Work

Studies show that despite the overall increasing number of disadvantaged learners in higher education, disadvantaged students are still under-represented in many countries, and their educational outcomes are not equal to other students (Bennett, Southgate, & Shah, 2016; Pitman et al., 2019; Harris et al., 2020). The growth of online learning in recent decades raises questions about access, equity and ethics (Zawacki-Richter, 2009). Online learning environments are a potential solution for widening access to education for diverse learners, such as students from low socioeconomic backgrounds, students with disabilities, and students from different regions. A qualitative study on online students followed by interviews with academic staff provided guidelines for providing an inclusive online learning environment (Stone, 2017). These guidelines emphasise the importance of understanding the nature and diversity of the online students, providing early intervention with students to engage and leveraging learning analytics to provide effective intervention and personalisation.

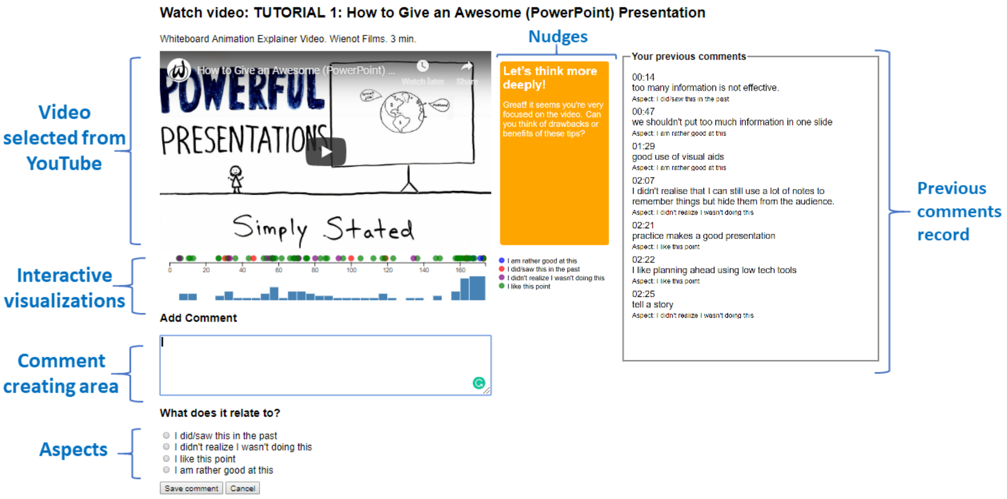
There have been several studies on under-represented students in online learning platforms. A recent study investigated the linguistic differences of self-reflections made by female and male students in an online chemistry class and analysed gender differences in the relationship between these linguistic features and learning outcomes (Lin, Yu, & Dowell, 2020). Another study investigated how a gameful LMS (Learning Management System) for five courses (sociology, communication, education, Honors program, movement science) at a large university have impacted the performance of under-represented students (Hayward, Schulz, & Fishman, 2021). This study showed that under-represented minority students were proportionally under-represented among A-earning students. Another study analysed the clickstream of students in an online chemistry course to investigate the self-regulated learning differences between first-generation students (who are the first in their family studying at the university) and traditional students (Rodriguez et al., 2021). The study revealed that first-generation college students classified as Early Planners (visited and watched the most videos when assigned) performed as well as their non-first-generation peers, but first-generation students in the Low Engagement group (had the lowest number of lecture video visits and a high number of watched videos closer to the due date) had the lowest average grades. This shows that minority students may benefit from utilising self-regulated learning strategies.

Language is identified as an important factor of dropout in MOOCs (Henderikx, Kreijns, & Kalz, 2018; Said, 2017). An early study on MOOCs reported language barriers non-native speakers experience, such as low reading speed, information overload and cognitive process and stress related to the visibility of their written responses (Sanchez-Gordon & Luján-Mora, 2014). Although delivering learning materials, providing translations and interface customisation in different languages have been successful in supporting EFL students (Colas, Sloep, & Garreta-Domingo, 2016; Murugesan, Nobes, & Wild, 2017; Navarrete & Luján-Mora, 2018), such approaches are time-consuming and require a lot of resources and effort (Lambert, 2020). According to the language barriers discussed earlier, simpler approaches were suggested for supporting EFL students, such as providing the ability to pause or regulate the video speed, giving access to a downloadable transcript or a specialised glossary and learning materials which use colour and visualisations (Sanchez-Gordon & Luján-Mora, 2015). However, adapting learning resources to the student's language competency requires a comprehensive investigation of characteristics of Native and EFL students, which is the focus of the work presented in this paper.

3. AVW-Space

AVW-Space aims at supporting engagement in two phases: 1) Personal space (Figure 1), where students watch videos and make comments, and 2) Social space (Figure 2), where students review and rate comments made by their peers. AVW-Space provides interactive visualisations of comments written by previous students in Personal space, which show the distribution of comments over various parts of the video. Students can also see the comment text when hovering over the comment (Mitrovic et al., 2019). Regarding supports suggested for EFL students, AVW-Space allows students to pause or restart videos, adjust the speed and enable auto-generated closed caption of the video. Additionally, comments shown to the whole class for rating in the Social space are anonymised to reduce the stress related to the visibility of comments.

AVW-Space provides nudges in the Personal space, which are personalised interventions to enhance constructive behavior. The nudges have been designed based on the results of initial studies with AVW-Space (Dimitrova et al., 2017; Mitrovic et al., 2019) to encourage students to write comments, use various aspects and improve the quality of comments. For instance, if a student is passively watching the video and has made no comments, the student will receive a Reminder nudge, stating that writing comments is beneficial for learning. Also, if the student used only one aspect in her/his comments, a Reminder nudge will be shown to draw the student's attention to other aspects. If a student often writes comments which only repeat the content of the video, he/she will receive a Quality nudge, asking the student to think about the causes/effect or cons/pros of the tips covered in the video. Also, a student who is watching the last part of the video and has made no self-reflective comments will be given a Quality nudge to think about his/her previous experience and plan for future improvements using the taught tips. Previous studies showed that Reminder nudges increased the number of comments (Mitrovic et al., 2019). A recent study on the Quality nudges also showed a significant improvement in the quality of comments (Mohammadhassan et al., 2021). However, these supports are generic, and more investigation on the differences of EFL/Native students is required to provide more specialised support for them.



Watch video: TUTORIAL 1: How to Give an Awesome (PowerPoint) Presentation

Whiteboard Animation Explainer Video. Wienot Films. 3 min.

Let's think more deeply!

Great! it seems you're very focused on the video. Can you think of drawbacks or benefits of these tips?

Your previous comments

00:14
too many information is not effective.
Aspect: I did/saw this in the past

00:47
we shouldn't put too much information in one slide
Aspect: I am rather good at this

01:29
good use of visual aids
Aspect: I am rather good at this

02:07
I didn't realise that I can still use a lot of notes to remember things but hide them from the audience.
Aspect: I didn't realize I wasn't doing this

02:21
practice makes a good presentation
Aspect: I like this point

02:22
I like planning ahead using low tech tools
Aspect: I like this point

02:25
tell a story
Aspect: I didn't realize I wasn't doing this

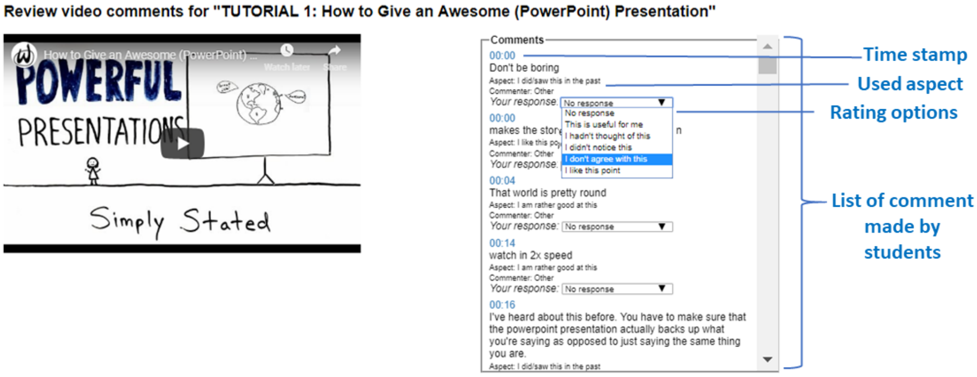
Add Comment

What does it relate to?

○ I did/saw this in the past
○ I didn't realize I wasn't doing this
○ I like this point
○ I am rather good at this

Nudges
Video selected from YouTube
Interactive visualizations
Comment creating area
Aspects
Previous comments record
How to Give an Awesome (PowerPoint) ...
Watch later
Share
POWERFUL
PRESENTATIONS
Simply Stated
0
20
40
60
80
100
120
140
160
● I am rather good at this
● I did/saw this in the past
● I didn't realize I wasn't doing this
● I like this point
Save comment
Cancel

Figure 1. A Screenshot of Personal Space.



Review video comments for "TUTORIAL 1: How to Give an Awesome (PowerPoint) Presentation"

Comments

00:00
Don't be boring
Aspect: I did/saw this in the past
Commenter: Other
Your response: No response ▼

00:00
makes the stor… n
Aspect: I like this po…
Commenter: Other
Your response:

00:04
That world is pretty round
Aspect: I am rather good at this
Commenter: Other
Your response: No response ▼

00:14
watch in 2x speed
Aspect: I am rather good at this
Commenter: Other
Your response: No response ▼

00:16
I've heard about this before. You have to make sure that the powerpoint presentation actually backs up what you're saying as opposed to just saying the same thing you are.
Aspect: I did/saw this in the past

Figure 2. A Screenshot of Social Space.

4. Materials and Methods

In order to compare EFL and Native students, we used the data from three previous studies with AVW-Space. Study 1 was conducted with an early version of AVW-Space with no nudges. Study 2 was conducted with the version providing only Reminder nudges, while in Study 3 both Reminder and Quality nudges were provided. The three studies were conducted in the same first-year engineering course at the University of Canterbury, in three consecutive years. In each study, students were invited to use the online training for presentation skills using AVW-Space to prepare for the presentation of their final project. The use of AVW-Space was voluntary. The participants were first invited to complete Survey 1, and were then instructed to watch and comment on the provided videos. There were four tutorial videos on giving presentations, and four examples of real presentations. Nudges were provided during this phase (in Studies 2 and 3). In the second phase, students were instructed to review and rate the comments made by their peers. Finally, Survey 2 was released to students.

Survey 1 contained demographic questions, questions about the participant's knowledge on giving presentations, experience and training in giving presentations, how often they used YouTube generally, and how often they used YouTube for learning. The last part of Survey 1 included the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich & de Groot, 1990). Survey 2 included the same questions about giving presentations to investigate whether students have increased their knowledge. Survey 2 also included two other questionnaires: NASA-TLX (Hart, 2006) to analyse the cognitive load of interacting with AVW-Space, and the Technology Acceptance Model (TAM) (Davis, 1989) to evaluate the perceived usefulness of AVW-Space. For knowledge questions in both surveys, the participants were asked to write everything they knew about visual aids, structure and delivery (one minute per question). The students' answers were marked automatically, using the ontology of presentation skills (Dimitrova et al., 2017). The marks for conceptual knowledge questions are used as the pre/post-test scores (CK1/CK2).

5. Results

There were 986 students (133 EFL and 853 Native) in all three studies who completed Survey 1. The first language of the majority of EFL students (44.92%) was Chinese and Vietnamese, while 17.39% spoke European languages (e.g. Dutch and Spanish) and 14.49% Indian languages (e.g. Hindi and Punjabi). Table 1 shows the distribution of EFL/Native students in the three studies and their CK1 scores. As Study 3 was conducted in 2020, there were fewer EFL students due to the COVID19 travel restrictions. We ran ANOVA on CK1 with study and Native/EFL as two factors. The test of between-subject effects showed that the study had no significant effect on CK1. However, whether students were Native speakers of English or EFL had a significant effect on CK1 (F = 40.44, p < .001); EFL students had significantly lower CK1 scores compared to Native students. As there are no differences in CK1 scores from the three studies, we combined all Native and EFL students, and report analyses done in the following sections.

Table 1. *Distribution of EFL/Native Students in Three Studies and Their CK1 (means and standard deviations in parentheses)*

	#All	#EFL	#Native	EFL CK1	Native CK1
Study 1	355	63	292	10.52 (6.48)	12.90 (5.62)
Study 2	338	42	296	8.19 (5.37)	13.55 (5.86)
Study 3	294	29	265	11.44 (6.39)	14.25 (5.38)
Total	986	133	853	9.99 (6.25)	13.54 (5.64)

5.1 Self-reported Learning Strategies

There were no significant differences between Native and EFL students on the scores for training/experience in giving presentations and using YouTube. Table 2 shows the scores on the MSLQ dimensions. We found significant differences in extrinsic goal orientation (t = 4.40, p < .001), rehearsal (t = 4.96, p < .001), self-regulation (t = 2.23, p < .01), organisation (t =2.46, p < 0.05) and critical thinking (t = 2.99, p < .001). Thus, EFL students reported strong meta-cognitive strategies, but reasons such as grades, rewards, evaluation by others, and competition motivate EFL students more than Native students.

Table 2. *MSLQ scores for EFL and Native Students, using a Likert scale from 1 (lowest) to 7 (highest).*

MSLQ Score	EFL	Native
Control of learning	5.65 (0.85)	5.61 (0.77)
Effort regulation	4.73 (1.04)	4.84 (0.98)
Elaboration	5.10 (1.06)	5.03 (0.89)
Extrinsic goal orientation***	5.80 (0.94)	5.38 (1.04)
Intrinsic goal orientation	5.28 (0.99)	5.12 (0.85)
Self-Regulation**	4.68 (0.92)	4.42 (0.70)
Organisation *	4.89 (1.10)	4.63 (1.12)
Rehearsal***	4.60 (1.16)	4.11 (1.09)
Self-efficacy	5.07 (0.96)	5.01 (0.93)
Task value	5.43 (0.95)	5.49 (0.80)
Critical thinking***	4.73 (1.13)	4.35 (1.07)

* p< .05, ** p < .01, *** p < .001

5.2 Engagement

To investigate students' engagement, we compared the number of comments students made, the number of videos they watched and the number of comments they rated (Table 3). We applied ANOVA on these activities, with study and EFL/Native as two fixed factors. The test of between-subject effects showed that the study had significant effect on the number of videos (F = 41.93, p < .001), comments (F = 5.49, p < .01) and ratings (F = 6.48, p < .01), due to the effect of nudges (Mitrovic et al., 2019, Mohammadhassan et al., 2021). However, EFL/Native only had a significant effect on the number of comments (F = 11.17, p < .05). Native students made significantly more comments than EFL students in Study 3 (F = 3.88, p < .05). In Study 3, which included Reminder and Quality nudges, the EFL students received significantly fewer nudges than Native students (F = 3.95, p < .05), but there was no significant difference in the number of nudges received by EFL/Native students in Study 2. The difference between EFL/Native students in Studies 2 and 3 could indicate that the model which triggers Quality nudges is tailored to the behavior of Native students more than EFL students.

Table 3. *Statistical Description of Activities for EFL and Native Students*

		Videos	Nudges	Comments	Ratings
Study 1	Native	6.86 (4.93)	None	4.13 (7.72)	3.79 (18.43)
	EFL	6.77 (4.52)	None	2.90 (6.97)	10.17 (46.25)
Study 2	Native	6.68 (3.96)	9.70 (10.30)	6.72 (10.26)	4.00 (19.05)
	EFL	7.40 (4.38)	9.43 (5.59)	5.38 (8.04)	1.16 (4.06)
Study 3	Native	7.03 (4.77)	20.38 (16.26)	11.43 (0.67)	21.75 (72.08)
	EFL	5.72 (3.90)	14.17 (12.71)	3.24 (2.02)	21.69 (85.91)

We also computed the linguistic and psychological features for comments using LIWC (Linguistic Inquiry Word Count) (Pennebaker et al., 2015, p. 201). LIWC takes a word count approach using dictionaries collected from various psychological constructs such as cognitive processes and perceptual process (Tausczik & Pennebaker, 2010). There were 273 comments made by 66 EFL students, and 2,318 comments made by 649 Native students on tutorial videos. After applying an independent t-test on LIWC features of the tutorial comments, we found no significant difference in the comment lengths, but the number of words per sentence in comments made by EFL students was significantly lower than for Native students. Table 4 shows the mean and standard deviation of LIWC features with significant differences for comments made by EFL and Native students on tutorial videos. EFL students used the first-person singular pronouns ("I", "my", etc.) and auxiliary verbs (such as "will" or "could") significantly less than Native students. Since comments showing self-reflection and self-regulation usually contain first-person pronouns (Gašević, Mirriahi, & Dawson, 2014; Jung & Wise, 2020), the differences in LIWC scores for EFL and Native students could mean that the Native students wrote more self-reflective and self-regulating comments. There were no significant differences in cognitive process features such as insight, certainty and differentiation. However, EFL students had a significantly higher score for causation (e.g. "because", "effect", etc.). There were no significant differences in the perceptual process such as seeing, feeling and hearing. However, comments made by EFL students had significantly higher positive emotion scores, as well as significantly lower scores for non-fluent words such as "hm" or "umm". In addition to LIWC features, we calculated the domain-specific ratios (Dimitrova et al., 2017) to measure how relevant comments are to the domain. The domain-specific ratio is the number of words from the domain ontology appearing in the comment, divided by the total number of words in the comment. The independent t-test on the domain-specific ratio showed no significant difference for the comments made by EFL and Native students on tutorial videos.

We also compared the linguistic features of comments made on example videos using independent t-test (Table 5). There were 149 comments made by 51 EFL students and 1,383 comments made by 307 Native students on example videos. Similar to tutorial comments, comments made by EFL students had significantly lower scores for non-fluent words. Also, comments made by EFL students lower score in using verbs and adverbs with present focus, but significantly higher domain-specific ratios. This could mean EFL students listed good practices of oral presentation rather than making complete sentences critiquing the presentation in the example video.

Table 4. S*ignificantly Different LIWC Features for Comments on Tutorial Videos*

LIWC Features	EFL	Native	Significance
First single pronoun	0.43 (1.93)	0.87 (3.15)	t = 3.22, p < .01
Auxiliary verbs	5.33 (8.17)	6.73 (9.14)	t = 2.64, p < .01
Causation	4.91 (8.11)	3.70 (6.95)	t = 2.35, p < .05
Positive emotions	9.97 (18.68)	7.50 (13.16)	t = 2.12, p < .05
Affiliation	1.78 (8.15)	0.74 (3.17)	t = 2.05, p < .05
Non-fluency	0.05 (0.55)	0.18 (1.75)	t = 2.61, p < .01
Word per Sentence	9.61 (7.81)	10.70 (8.40)	t = 2.06, p < .05

Table 5. S*ignificantly Different LIWC Features for Comments on Example Videos*

LIWC Features	EFL	Native	Significance
Focus present	8.59 (10.27)	10.99 (11.03)	t = 2.53, p < .05
Non-fluency	0.16 (0.75)	0.87(4.51)	t = 5.18, p < .01
Domain-specific proportion	0.34 (0.26)	0.29 (0.24)	t = 2.14, p < .05
Unique Domain-specific proportion	0.34 (0.26)	0.30 (0.24)	t = 2.11, p < .05

We classified students post-hoc into different categories based on ICAP framework (Chi & Wylie, 2014). The ICAP framework classifies learners' overt behaviours into four categories: Interactive, Constructive, Active and Passive. Passive learners receive information by only watching videos. Active students perform additional actions like note-taking, but their annotations merely repeat the received information. Constructive learners add new information that was not explicitly taught, by reflecting on their knowledge or making inferences. The Interactive category is not applicable in our research since AVW-Space does not support direct interaction between students. We labelled students as Constructive if they had at least three comments showing critical thinking, self-reflection or self-regulation (based on the median number of such comments). Students who had less than three comments of that type were labelled as Active. Finally, students who only watched the videos and did not make any comments were classified as Passive.

Table 6 shows the distribution of EFL/Native students in the Passive, Active and Constructive categories. A chi-square test of homogeneity revealed a significant difference (Chi-square = 16.76, p < .001), with the effect size (Phi) of .13 (p< .001) on all three studies (the *Overall* column in Table 6). We applied a post-hoc analysis to compare different categories using the z-test with a Bonferroni correction. For EFL students, the proportion of the Constructive category was significantly lower (p <

.05) than other categories, while for Native students, the proportions of the different categories were similar. It can be seen that the majority of EFL students were passive, which indicates the need to provide more focused support for them.

Table 6. *Distribution of EFL/Native Students in ICAP Categories for Different Studies*

		Study 1	Study 2	Study 3	Overall
EFL	Passive	34 (54.0%)	19 (45.2%)	15 (51.7%)	68 (50.7%)
	Active	26 (41.3%)	15 (35.7%)	9 (31.0%)	50 (37.3%)
	Constructive	3 (4.8%)	8 (19.0%)	5 (17.2%)	16 (11.9%)
Native	Passive	161 (55.1%)	108 (36.5%)	60 (22.6%)	329 (38.6%)
	Active	95 (32.5%)	107 (36.1%)	105 (39.6%)	307 (36.0%)
	Constructive	36 (12.3%)	81 (27.4%)	100 (37.7%)	217 (25.4%)

We investigated the effect of nudges on EFL/Native students' engagement. A chi-square test of homogeneity between the studies and the ICAP categories of EFL students revealed a significant difference (Chi-square = 6.16, $p < .05$) with the effect size (Phi) = .21 ($p < .05$). Adding Reminder nudges in Study 2 raised the percentage of constructive EFL students significantly compared to Study 1 ($p < .05$), but the percentage of constructive EFL students was not significantly different between Studies 2 and 3. Also, there was no significant difference in the proportion of passive EFL students between the studies.

We also applied a chi-square test of homogeneity between the studies and ICAP categories for Native students, which showed a significant difference (Chi-square = 76.39, $p < .001$) with effect size (Phi) of .29 ($p < .001$). The percentage of constructive students increased significantly by including the Quality nudges in Study 3, compared to Study 1 with no nudges and Study 2 with only Reminder nudges. Unlike EFL students, the percentage of passive students decreased significantly by adding the Reminder nudges (Study 2) and Quality nudges (Study 3). Thus, the nudges were more effective for Native students than EFL students.

5.3 Learning

We compared the CK2 of EFL and Native students to find out whether there was a difference in learning. Since only 622 of students completed Survey 2, we only have CK1 and CK2 for 80 EFL students and 542 Native students. We ran an ANCOVA on CK2 scores, with CK1 as a co-variate, and study and being EFL/Native as two fixed factors. We found no significant difference in learning between different studies. After applying the mean adjustment on CK2 using Bonferroni, we found that EFL students learned significantly less (12.17 ± .64) than Native students (14.37 ± .23); (F = 10.37, $p < .001$). However, lower CK1 and CK2 scores in EFL students could be due to language barriers that EFL students might struggle with in answering the conceptual knowledge questions in Surveys 1 and 2, while they might have learnt the skill.

In order to find the factors influencing learning for EFL/Native students, we ran a generalised linear regression using CK1 and the number of comments made to predict CK2, with being EFL/Native as the fixed factor. The models fitted with Akaike's Information Criterion (AIC) = 3,793.12 (Table 7). CK1 and the number of comments were significant predictors. Each additional point on CK1 has a 0.15 extra effect on CK2 for Native students (the interaction effect of CK1 *Native is 0.14).

Table 7. *Significant Predictors of CK2 for EFL/Native Students*

Variables	Coefficient	Significance
Intercept	4.96	$p < .001$
CK1	0.50	$p < .001$
CK1*Native	0.15	$p < .005$
Comment	0.21	$p < .001$
CK1*Comment	-.007	$p < .05$

5.4 Subjective Opinions

We investigated the responses of EFL/Native students to the NASA-TLX and TAM questionnaires. TLX-NASA uses a Likert scale from 1 (lowest) to 20 (highest), and TAM uses a Likert scale from 1 (highest) to 7 (lowest). In all studies, NASA-TLX questions asked participants to report their perceived cognitive load during commenting on videos and rating comments written by their peers. There was no significant difference in the perceived mental demand, required effort and confidence in performance for the two tasks between Native/EFL students. However, EFL students found the rating task significantly more frustrating (8.81 ± .65) than Native students (7.30 ± .24); (F = 4.70, p < .05). Also EFL students perceived frustration during commenting (8.85 ± .66) significantly more than Native students (7.43 ± .24); (F = 4.08, p < .05). We also found that students had no significantly different opinions on the usefulness of AVW-Space based on the responses to the TAM questionnaire, except that the EFL students had lower scores (3.55± .19) for "I think I would like to use AVW-Space frequently" (F = 9.65, p < .01) and "*If I am provided the opportunity, I would continue to use AVW-Space for informal learning*" (3.50 ± .19, F = 5.60, p < .05) compared to Native students (4.17 ± .07 and 4.00 ± .07, respectively).

We also looked at the feedback EFL students provided on nudges. Some EFL students reported that the nudges distracted them from videos (e.g. *"not very useful, took away from the video"*) or they were not confident in writing comments (e.g. *"I am not confident"*, "*They were not useful since I did not know what to do to start with"*). Also, some responses from passive EFL students showed that they misunderstood the purpose of nudges, such as: *"[nudges helped me] to understand some features I did not know*". There was also some positive feedback from passive EFL students, reporting nudges were useful (e.g. "*Give me directions*", "*Somewhat helpful to remind the user to write a comment*"). However, given that these students were in the Passive category, the nudges were not effective enough for these students to encourage them to make comments.

6. Discussion and Conclusions

We investigated the differences between EFL and Native students in their learning strategies, engagement and learning outcomes in AVW-Space. We found that majority of EFL students watched educational videos without writing comments. Furthermore, EFL students had lower conceptual knowledge scores before and after the study in comparison to Native students. Although adding Reminder nudges increased constructive behaviour in EFL students, including Quality nudges was not effective for EFL students compared to Native students. Linguistic analysis of comments showed significantly fewer indicators of self-reflection in comments made by EFL students than Native students. The linear regression revealed the importance of commenting for EFL/Native students. The comparison of subjective opinions of the EFL student showed confusion about nudges, lack of confidence in making comments and frustration in commenting and reviewing task. Therefore, more focused support should be provided to EFL students to help them benefit from VBL as much as Native students.

The identified significant differences between these two categories of students allow for specifying tailored support for EFL students. Comparing learning strategies of EFL and Native students showed that EFL students are more oriented towards extrinsic goals. One way to increase their motivation to write comments is to provide a dashboard which visualises their progress or allows them to compare themselves with the class, since visualisation has been effective in increasing motivation in various learning activities (Aguilar et al., 2021). We found that EFL students reported stronger metacognitive strategies than Native students. However, only a minority of them wrote comments showing self-reflection and self-regulation. Therefore, including more self-regulatory activities, such as goal setting and monitoring previously received nudges and written comments, could activate self-regulation in EFL learners.

Providing downloadable transcripts for videos could be helpful to EFL students, as suggested in the literature. Additionally, showing each nudge for a longer period of time could be helpful to EFL students to comprehend nudges. Adding a glossary of main concepts could also help EFL students in understanding videos and improving their vocabulary. Finally, providing information and feedback using colours and signs could also reduce the cognitive and information load for EFL students.

The main limitation of this research was the context of this study (oral presentation skills), since the nature of this domain involves language proficiency. Thus, applying similar analysis in the contexts of technical skills such as programming could result in different insights. Another limitation of this research is the low percentage of EFL students in the study population. Also, the analysis of the learning outcomes of EFL/Native students in this research was only based on the students' written responses to the conceptual knowledge questions, which again require English competency. Therefore, a more sophisticated approach is required to assess the students' presentation skills before and after using the system, regardless of their English competency.

This research contributes to understanding the requirements for improving inclusiveness in computer-assisted learning environments and improving equity in the learning experience and outcomes for non-native English speakers. Although this research focused only on a particular video-based learning platform, the important findings obtained in this research encourage researchers to investigate the equity for non-native English speakers in other platforms and propose effective approaches to achieve this goal.

References

Aguilar, S.J., Karabenick, S.A., Teasley, S.D., Baek, C., (2021). Associations Between Learning Analytics Dashboard Exposure and Motivation and Self-regulated Learning. *Computers & Education, 162, 104085.*

Bennett, A., Southgate, E., & Shah, M. (2016). *Chapter 15—Global Perspectives on Widening Participation: Approaches and Concepts*. Chandos Publishing.

Chi, M. T. H., & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist, 49*(4), 219–243.

Colas, J.-F., Sloep, P. B., & Garreta-Domingo, M. (2016). The Effect of Multilingual Facilitation on Active Participation in MOOCs. *The International Review of Research in Open and Distributed Learning*, *17*(4), 280–314.

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319–340. JSTOR. https://doi.org/10.2307/249008

Dimitrova, V., Mitrovic, A., Piotrkowicz, A., Lau, L., & Weerasinghe, A. (2017). Using Learning Analytics to Devise Interactive Personalised Nudges for Active Video Watching. *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pp. 22–31. Bratislava, Slovakia: ACM.

Gašević, D., Mirriahi, N., & Dawson, S. (2014). Analytics of the Effects of Video Use and Instruction to Support Reflective Learning. *Proceedings of the 4th International Conference on Learning Analytics And Knowledge*, 123–132. New York, NY, USA: ACM.

Giannakos, M. N., Sampson, D. G., & Kidziński, Ł. (2016). Introduction to smart learning analytics: Foundations and developments in video-based learning. *Smart Learning Environments*, *3*(1), 1–9.

Harris, R. B., Mack, M. R., Bryant, J., Theobald, E. J., & Freeman, S. (2020). Reducing achievement gaps in undergraduate general chemistry could lift underrepresented students into a "hyperpersistent zone". *Science Advances*, *6*(24), eaaz5687.

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908.

Hayward, C., Schulz, K., & Fishman, B. (2021). Who Wins, Who Learns? Exploring Gameful Pedagogy as a Technique to Support Student Differences. *Proc. 11th International Learning Analytics and Knowledge Conference*, 559–564. New York, NY, USA: Association for Computing Machinery.

Henderikx, M., Kreijns, K., & Kalz, M. (2018). A Classification of Barriers that Influence Intention Achievement in MOOCs. In V. Pammer-Schindler, M. Pérez-Sanagustín, H. Drachsler, R. Elferink, & M. Scheffel (Eds.), *Lifelong Technology-Enhanced Learning* (pp. 3–15). Cham: Springer International Publishing.

Jung, Y., & Wise, A. F. (2020). How and How Well Do Students Reflect? Multi-Dimensional Automated Reflection Assessment in Health Professions Education. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 595–604. New York, NY, USA: Association for Computing Machinery.

Lambert, S. R. (2020). Do MOOCs contribute to student equity and social inclusion? A systematic review 2014–18. *Computers & Education*, *145*, 103693.

Lin, Y., Yu, R., & Dowell, N. (2020). LIWCs the Same, Not the Same: Gendered Linguistic Signals of Performance and Experience in Online STEM Courses. In I. I. Bittencourt, M. Cukurova, K. Muldner, R.

Luckin, & E. Millán (Eds.), *Artificial Intelligence in Education* (pp. 333–345). Cham: Springer International Publishing.

Mitrovic, A., Dimitrova, V., Lau, L., Weerasinghe, A., & Mathews, M. (2017). Supporting Constructive Video-Based Learning: Requirements Elicitation from Exploratory Studies. In E. André, R. Baker, X. Hu, Ma. M. T. Rodrigo, & B. du Boulay (Eds.), *Artificial Intelligence in Education* (pp. 224–237). Springer International Publishing.

Mitrovic, A., Dimitrova, V., Weerasinghe, A., & Lau, L. (2016). Reflective Experiential Learning: Using Active Video Watching for Soft Skills Training. *Proceedings of the 24th International Conference on Computers in Education*, 192–201. Asia-Pacific Society for Computers in Education.

Mitrovic, A., Gordon, M., Piotrkowicz, A., & Dimitrova, V. (2019). Investigating the Effect of Adding Nudges to Increase Engagement in Active Video Watching. In: S. Isotani et al. (Eds.) Proc. 20th Int. Conf. AIED 2019, LNAI 11625, pp. 320-332, Springer Nature Switzerland.

Mohammadhassan, N., Mitrovic, A., Neshatian, K., Dunn, J. (2020) Automatic quality assessment of comments in active video watching using machine learning techniques. In: So, H.J. et al. (Eds.) Proceedings of the 28th International Conference on Computers in Education, pp. 1-10. Asia-Pacific Society for Computers in Education.

Mohammadhassan, N., Mitrovic, A., Neshatian, K., & Dunn, J. (2021). Investigating the Effect of Nudges for Improving Comment Quality in Active Video Watching (under review).

Murugesan, R., Nobes, A., & Wild, J. (2017). A MOOC approach for training researchers in developing countries. *Open Praxis*, *9*(1), 45–57.

Navarrete, R., & Luján-Mora, S. (2018). Bridging the accessibility gap in Open Educational Resources. *Universal Access in the Information Society*, *17*(4), 755–774.

Pennebaker, J., Boyd, R., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Pintrich, P. R., & de Groot, E. V. (1990). Motivational and Self-Regulated Learning Components of Classroom Academic Performance. *Journal of Educational Psychology*, *82*, 33–40.

Pitman, T., Roberts, L., Bennett, D., & Richardson, S. (2019). An Australian study of graduate outcomes for disadvantaged students. *Journal of Further and Higher Education*, *43*(1), 45–57.

Rodriguez, F., Lee, H. R., Rutherford, T., Fischer, C., Potma, E., & Warschauer, M. (2021). Using Clickstream Data Mining Techniques to Understand and Support First-Generation College Students in an Online Chemistry Course. *Proc. 11th International Learning Analytics and Knowledge Conference*, 313–322. New York, NY, USA: Association for Computing Machinery.

Said, G. R. E. (2017). Understanding How Learners Use Massive Open Online Courses and Why They Drop Out: Thematic Analysis of an Interview Study in a Developing Country. *Journal of Educational Computing Research*, *55*(5), 724–752.

Sanchez-Gordon, S., & Luján-Mora, S. (2015). Accessible blended learning for non-native speakers using MOOCs. *International Conference on Interactive Collaborative and Blended Learning*, pp. 19–24.

Sanchez-Gordon, S., & Luján-Mora, S. (2014). MOOCs Gone Wild. *Proceedings of the 8th International Technology Education and Development Conference.*, 1449–1458.

Stone, C. (2017). *Opportunity through online learning: Improving student access, participation and success in higher education* (pp. 26–48). Perth: The National Centre for Student Equity in Higher Education (NCSEHE). Curtin University.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.

Zawacki-Richter, O., Bäcker, E.M., Vogt, S. (2009). Review of Distance Education Research (2000 to 2008): Analysis of Research Areas, Methods, and Authorship Patterns. *International Review of Research in Open and Distributed Learning*, *10*(6), 21–50.

Improving Knowledge Tracing through Embedding based on Metapath

Chong JIANG[a], Wenbin GAN[b], Guiping SU[a], Yuan SUN[b] & Yi SUN[a*]
[a]*University of Chinese Academy of Science, China*
[b]*National Institute of Informatics, Japan*
*sunyi@ucas.ac.cn

Abstract: The goal of knowledge tracing (KT) is to track students' knowledge status and predict their future performance based on their learning logs. Although many researches have been devoted to exploiting the input information, they do not strictly distinguish between questions and the involved skills when taking the learning logs as input, and hence leading to performance degradation due to the fact that the inherent relations between skills and questions are not fully utilized. To solve this issue, we propose an embedding pre-training method based on metapath by explicitly considering the relations between skills and questions in the domain. Specifically, we construct a heterogeneous graph composed of skills and questions, and obtain the meaningful embeddings of nodes using the metapath2vec method, hence the explicit relation information can be embedded in the dense representation of skills and questions while still maintaining their own characteristics. Adopting these pre-trained embeddings to existing models, experiments on three public real-world datasets demonstrate that our method achieves the new state-of-the-art performance, with at least 1% absolute AUC improvement.

Keywords: Knowledge Tracing, matapath, network embedding

1. Introduction

Knowledge tracing is a fundamental task in intelligent tutoring systems to provide adaptive services to learners. The main aim of KT is to track students' knowledge status based on their learning records over time and predict their future performance accordingly.

At present, considerable progress has been achieved on this task due to the prevalence of online education. The existing models can be divided into two categories: traditional method and deep learning-based method. Bayesian Knowledge Tracing (BKT) (Corbett & Anderson, 1994) is a typical model in traditional method, which infers the evolution of a student's skill mastery using a Markov model. Deep Knowledge Tracing (DKT) (Piech, Spencer, & Sohl-Dickstein, 2015) is the first model to apply deep learning method into KT task, and has made a breakthrough.

For the KT task, the number of questions in a tutoring system generally far exceeds the number of skills. A skill may relate to many questions and a question may also correspond to more than one skill. To simplify the modeling process and improve the prediction efficiency, KT task deals with the model input in a unified way: KT models in both the traditional and deep learning categories conduct the KT process and make predictions based on the skills instead of the questions themselves. They assume that skill mastery can potentially reflect the possibility that a student can correctly solve a question incorporating that skill to some extent, hence the input to the KT models is actually skill tags alternatively. Moreover, each question is assumed to associated with only one skill. For a question containing multiple skills, a new skill is generated to represent the combination of the multiple skills.

Therefore, the relations between skills and questions, and their own characteristics are neglected, which can cause two important issues: the first one is that the tracking of students' knowledge states only stays at skill-level, and cannot truly reflect the actual ability of students to correctly solve the related questions; the second one is that the neglected relation and characteristic information are essential in predicting students' performance, and tracing students' knowledge states without considering this information will potentially cause the performance degradation. As shown in Figure 1, question q_1 is related to skill s_1, question q_2 is related to skill s_1 and skill s_2, question q_3 and question

q_4 share the same skill s_2. The main deep models track students' states at skill-level, when taking the input to the models, the question q_1 is replaced by skill s_1, the question q_3 and question q_4 are both replaced by skill s_2 though they have different difficulties. As for the question q_2, it is usually replaced by a new skill tag s_3, which means the combination of skill s_1 and skill s_2. Although the existing models perform well in the skill-level prediction, the problem of incorporating the more informative embedding representation by considering the relations between skills and questions, and their own characteristics, into the KT models to achieve more precise prediction of student performance still remains under-explored.

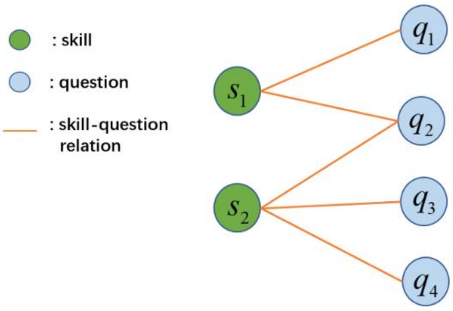


Figure 1. A Simple Example of Question-skill Relation.

In this paper, we take a further step towards exploiting the relations between skills and questions and obtaining the meaningful embedding of questions and skills together. Inspired by PEBG method (Liu, Y., Yang, Y & Yu, 2020), which learns a low-dimensional embedding for each question on the side information and the side information includes question difficulty and three kinds of relations contained in a bipartite graph between questions and skills, we abstract the relations between questions and skills as a heterogeneous bipartite graph considering that a skill may relates to many questions and a question may corresponds to more than one skill, and construct a series of meta-path to sample the whole graph. Deep embedding representation of each node in the graph is then obtained by not only using the own information, but also aggregating the information from connected skills and questions by utilizing the relations of skill-question in the graph. Finally, we incorporate the pre-training embeddings with the existing deep KT models to test its performance.

The contributions of this paper can be summarized as follows:

- By training the embeddings of questions and skills, we can track students' knowledge states at both skill- and question-level.
- We propose a pre-training approach to learn the pre-training embedding of skills and questions by leveraging the explicit relations between skills and questions in the heterogeneous bipartite graph, and incorporate the embeddings into the KT process.
- The extensive experiments on three benchmark datasets show that our model outperforms the existing state-of-the-art models, with at least 1% absolute AUC improvement.

2. Related Work

2.1 Knowledge Tracing

Existing KT models can be divided into two categories: traditional models and deep models, and deep models have shown overwhelmingly better performance than the traditional models. In this paper, we only focus on the deep KT models. Deep knowledge tracking (DKT) model (Piech, Spencer, & Sohl-Dickstein, 2015) is a milestone work to apply deep learning model in KT task, it uses a recurrent neural network (RNN) to model learners' learning process by taking the one-hot embedding representation of exercises and interactions. The other researchers subsequently have proposed various models to improve the DKT, the improvements are mainly in two aspects.

The first category improves the DKT model by designing different kinds of neural network structures to model students' learning process. Dynamic Key Value Memory Networks (DKVMN) (Zhang & Yeung, 2017) introduces Memory-Augmented Neural Networks (MANN) to solve the KT task and abstractly extracts the knowledge states of students. Graph-Based Knowledge Tracing (GKT) (Nakagawa & Matsuo, 2019) introduces the graph neural network (GNN) to solve the KT task which reformulates the KT task as a time series node-level classification problem in GNN.

The second category improves the DKT model by taking different embedding representations as input. The one-hot embedding representation of DKT only contains extremely sparse skill information and neglects the special characteristics of problems. Moreover, this representation can't reflect the relationships between skill and question. To solve these issues, Exercise-Aware Knowledge Tracing (EKT) (Liu, Q & Hu, 2019) trains the students' exercise records and the corresponding text content into a new embedding as the input of the network. Context-Aware Attentive Knowledge Tracing (AKT) (Ghosh & Lan, 2020) uses the transformer to train the embeddings of questions and skills based on the prior knowledge. Deep Knowledge with Convolutions (CKT) (Yang & Lu, 2020) uses three-dimensional convolution to aggregate the answer information of students in a period of time, and then trains the corresponding embedding as the input of LSTM network. KTM-DLF (Gan & Sun, 2020) model students' learning and forgetting behaviors by taking account of their memory decay and the benefits of attempts when an item can involve multiple KCs.

In addition to the above progress, the application of graph neural network and dense embedding to KT task also achieves good results. GIKT (Yang & Yu, 2020) uses a Graph Convolutional Network to aggregate the relationship between question and skill and obtain the skill embeddings as final input to RNN models. Following this line, the PEBG proposes a pre-training framework to get the embeddings of question. This paper also applies graph neural network and pre-trained dense embedding in KT task, however, different from these models, it considers the relations between skills and questions as a heterogeneous information graph and uses the metapath based method to learning the dense embeddings of skills and problems, and hence it can track students' knowledge states at both skill- and question-level, and obtains better performance.

2.2 Network Embedding

Network embedding (Chen & Skiena, 2018) aims to learn the low dimensional potential representation of nodes in the network. The learned feature representation can be used as the feature of various graph-based tasks, such as classification, clustering, link prediction and visualization. DeepWalk (Perozzi & Skiena, 2014) was proposed as the first network embedding method using deep learning. DeepWalk treats nodes in graph as words and generating short random walks as sentences. Then, neural language models such as Skip-gram (Mikolov & Dean, 2013) can be applied on these random walks to obtain network embedding. On the basis of DeepWalk, node2vec (Grover & Leskovec, 2016) adopts a biased wandering strategy which contains DFS and BFS. The network embedding method mentioned above is mainly used in isomorphic graphs. For heterogeneous graphs, it has different node types and edge types. Metapath2vec (Dong, Chawla & Swami, 2017) uses random walks based on meta path to construct the heterogeneous neighborhood of each node, and then uses skip gram model to complete the node embedding.

3. Problem Definition

3.1 Definition 1(Knowledge Tracing)

The KT task can be formally defined as follows: given a student's exercising records over a period of time $X_t = (x_1,x_2,\ldots,x_t)$, the KT model predicts the student's performance at the next moment x_{t+1}. Each interaction is composed of a question and the label indicating the correctness of student answer, hence x_t can be expressed as a pair of (q_t,a_t). The ordered pair indicates that the student has answered question q_t at time t and the correctness is a_t. The KT task aims to predict the probability $P(a_{t+1}{=}1|q_{t+1},X_t)$ of the student answering question q_{t+1} correctly.

3.2 Definition 2(Question-Skill Relation Graph)

A skill s_i may relate to many questions such as $\{q_1,\ldots,q_m\}$ and a question q_i may contain many skills such as $\{s_1,\ldots,s_n\}$. Here, we abstract a heterogeneous bipartite graph $G=(S,Q,R)$, where $R=[r_{ij}]\in\{0,1\}^{|S|\times|Q|}$ is a binary adjacency matrix and S means the set of all skills and the Q means the set of all questions. If question q_j contains skill s_i , there is edge between them in the graph G, and the entry in adjacency matrix $r_{ij}=1$; otherwise $r_{ij}=0$. In our model, the edge is bidirectional as shown in Figure 2.

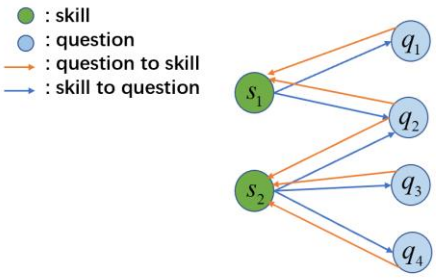


Figure 2. The Bidirectional Question-skill Relation Graph.

3.3 Definition 3(Meta-path)

Meta-path is a path containing a sequence of relations defined between objects of different types. The specific form is

$$A_1 \overset{R_1}{\to} A_2 \overset{R_2}{\to} \ldots \overset{R_l}{\to} A_{l+1} \tag{1}$$

It represents a compound relation between node types A_1 and A_{l+1} .The probability that the walker will take one step specified in the path is:

$$P(v_{t+1}|v_t,\rho)=\begin{cases}\dfrac{1}{|N_{t+1}(v_t)|}, & (v_t,v_{t+1})\in E, \phi(v_{t+1})=A_{t+1}\\ 0, & otherwise\end{cases} \tag{2}$$

Where $N_t(v)$ denotes v 's neighborhood with the t^{th} type of nodes and $\phi(v_{t+1})$ indicates the node v_{t+1} is in type of A_{l+1}. The constructed Meta-path is usually required to be symmetric based on the types to facilitate expansion. For a given Graph $G=(V,E,T)$ and a node v, we learn the embedding by maximizing the probability:

$$\arg\max_{\theta} \sum_{v\in V}\sum_{t\in T_V}\sum_{c_t\in N_t(v)} \log p(c_t \mid v;\theta) \tag{3}$$

where c_t is the context of node v and $p(c_t|v;\theta)$ is commonly defined as a Softmax function. In our question-skill heterogeneous bipartite graph, we only have two node types and two relation types.

4. Method

This section introduces our proposed method. Metapath2vec is applied to learn question and skill embeddings aggregated on the question-skill relation graph, and then a recurrent layer is used to model the sequential change of knowledge state. Then, we design an interaction module for the final prediction. The process of learning meta-path embedding representation of skill and question is shown in Figure 3. The recurrent layer and the interaction module for modeling student performance are shown in Figure 4.

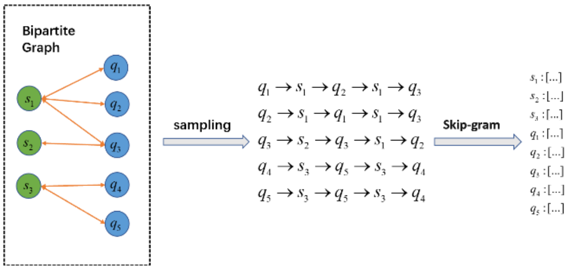

Bipartite Graph
s_1
s_2
s_3
q_1
q_2
q_3
q_4
q_5
sampling
$q_1 \to s_1 \to q_2 \to s_1 \to q_3$
$q_2 \to s_1 \to q_1 \to s_1 \to q_3$
$q_3 \to s_2 \to q_3 \to s_1 \to q_2$
$q_4 \to s_3 \to q_5 \to s_3 \to q_4$
$q_5 \to s_3 \to q_5 \to s_3 \to q_4$
Skip-gram
s_1 :[...]
s_2 :[...]
s_3 :[...]
q_1 :[...]
q_2 :[...]
q_3 :[...]
q_4 :[...]
q_5 :[...]

Figure 3. The process of Metapath based Embeddings.

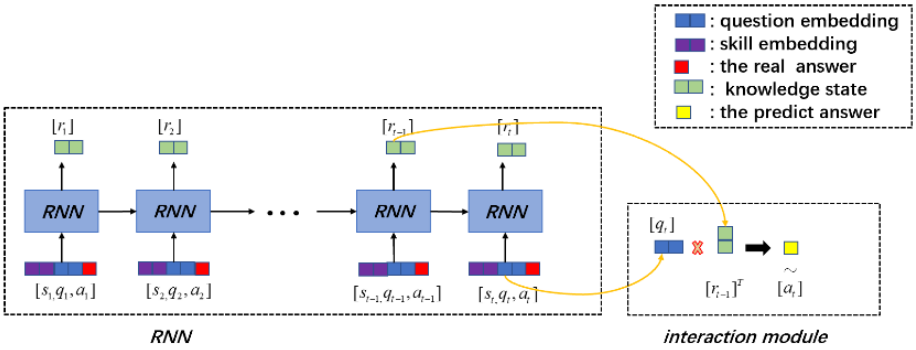


Figure 4. The Recurrent Layer and the Interaction Module.

4.1 Pre-training Question and Skill Embedding

As shown in Figure 3, a heterogeneous bipartite graph is constructed based on the relationship between skill and question in the dataset. We use a two-way arrow to show the connection between skill and question. These arrows represent the “question to skill” and “skill to question” relations in Figure 2. In the following, we use send to represents “question to skill” and back to represents “skill to question”. To predict whether a student can answer a question correctly, we need to understand the correlation between the question and the student’s knowledge state. This requires a careful consideration of the relationship between question and skill when coding them.

After constructing the relation graph of question and skill, a sampling process is conducted in the heterogeneous bipartite graph based on Metapath. Here we define a Metapath with five nodes, which is defined as follows:

$$Q \xrightarrow{send} S \xrightarrow{back} Q \xrightarrow{send} S \xrightarrow{back} Q \tag{4}$$

where Q means “question” and S means “skill”. By constructing such Metapath, we can aggregate “question”, “skill” and their relationship in a path at the same time. When we sample the path from the graph, we also set condition: the number of sampling paths must be equal to the number of questions, which requires that the starting node of each path is a different question node. As shown in Figure 3, there are five question nodes in the figure, and we get five paths by sampling. The advantage of setting this condition is that we can start from each question node when sampling, hence we can collect the information from the whole graph.

Finally, the unified d dimension embeddings of all the questions and skills can be obtained by training the sampled paths using metapath2vec (Dong, Chawla & Swami, 2017).

4.2 Metapath-Based Knowledge Tracing

We use DKT model as the baseline model and extend it with the pre-trained embeddings of questions and skills to formulate our model. In our model, we use combination of skill embedding and question embedding to replace one-hot embedding in DKT to trace students’ knowledge state and use question embedding to directly predict the probability of students correctly answering the next question.

In order to model the learning process of students, we use the RNN to process the exercise sequence, as shown in Figure 4. The input features x_t to our model is the combination of the skill embeddings and question embeddings attached with the students' answer at time t, hence the dimension of final input features is $2{*}d+1$. In addition, we can also use skill embeddings or question embeddings alone to track students' knowledge status, and the dimension of input features is $d+1$.

$$x_t=[s_t,q_t,a_t] \tag{5}$$

The formula of knowledge tracing process in RNN is as follows:

$$h_t=\tanh(W_{hx}x_t+W_{hh}h_{t-1}+b_h) \tag{6}$$

$$r_t=\sigma(W_{yh}h_t+b_y) \tag{7}$$

Where the x_t ,h_t ,r_t represents the input features, the hidden state and output state. The output state r_t represents a students' current knowledge state. The prediction result on the correctness of the student on the next question is obtained by interacting the output state in the current step with the question embedding at the next moment.

$$\tilde{a}_t=q_t\odot r_{t-1}^T \tag{8}$$

To optimize our model, we update the parameters in our model using gradient descent by minimizing the Cross Entropy Loss between the predicted probability of answering correctly and the true label of the student's answer:

$$\mathcal{L}=-\sum_t(a_t\log\tilde{a}_t+(1-a_t)\log(1-\tilde{a}_t)) \tag{9}$$

5. Experiments

We conduct several experiments to investigate the performance of our model. We first compare the performance of our model with five baselines on three public datasets. Then we conduct some ablation studies to investigate the effectiveness of our proposed model.

5.1 Datasets

We conduct experiments on three public datasets: ASSISTments2009, ASSISTments2012 and EdNet. The statistical information of these datasets is reported in Table 1.

Table 1. *Dataset Statistics*

	ASSIST09	ASSIST12	EdNet
#students	3,852	27,485	5,000
#questions	17,737	53,065	12,150
#skills	167	265	1774
#exercises	282,619	2,709,436	676,974
question pre skill	107	200	6.849
skill pre question	1.197	1.000	2.280
attempts per question	15	51	56
attempts per skill	1,692	10,224	381

ASSISTments2009 was collected from ASSISTments online education platform during the school year 2009-2010. For this dataset, we remove records without skills and scaffolding problems. Similar to other knowledge tracking methods, we also remove users with less than three records from the original dataset. In ASSISTments2009 dataset, it has 3852 students with 167 skills, 17,737 questions and 282,619 exercises.

ASSISTments2012 was collected from ASSISTments online education platform during the school year 2012-2013. ASSISTments2012 is different from ASSISTments2009 in that it has one characteristic that needs special emphasis. In ASSISTments2012 dataset, each question is only related to one skill, but one skill still corresponds to several questions. After the same data processing as ASSIST09, it has 2,709,436 exercises with 27,485 students, 265 skills and 53,065 questions.

EdNet was collected by (Choi, Lee & Heo,2020). In EdNet dataset, we only choose part of it. We choose EdNet-KT1 which has 676,974 exercises with 5,000 students, 1,774 skills and 12,150 questions after consistent processing with other data sets.

5.2 Baselines

We compare with five models, as follows:

BKT (Corbett & Anderson,1994) is a Bayesian model defined by initial knowledge, learning rate, slip and guess parameters. It models the knowledge state of the skill as a binary variable.

DKT (Piech, Spencer & Sohl-Dickstein, 2015) is the first method that introduces deep learning into knowledge tracing task. It uses recurrent neural network to model knowledge state of students.

DKVMN (Zhang, Shi & Yeung, 2017) introduces Memory-Augmented Neural Networks (MANN) to solve the KT task and abstractly extracts the knowledge states of students.

PEBG (Liu, Yang & Yu, 2020) uses the bipartite graph of question-skill relations to obtain question embeddings, which provides plentiful relation information.

AKTHE (Zhang, Du & Sun, 2020) capture the relevance of historical data to the current state by using attention mechanism.

5.3 Implementation Details

We implement all the compared methods using Pytorch. In the Metapath2vec, we use heterogeneous skip-gram model to get the embeddings and set the window size as 5, the batch as 128, the number of epoch as 10, the learning rate as 0.01. The Pre-traing embeddings of skill and question are both 10 dimensions.

In the implementation of RNN, a RNN with only one hidden layer is used and the max length of the RNN is set to 50. The batch size is set to 64 and the max epoch is 50, the learning rate is set to 0.001.

Table 2. *The AUC Results over Three Datasets*

	ASSIST09	ASSIST12	EdNet
BKT	0.6571	0.6204	0.6027
DKT	0.7561	0.7286	0.6822
DKVMN	0.7550	0.7283	0.6967
PEBG+DKT	0.8287	0.7665	0.7765
AKTHE	0.8310	0.7782	0.7757
OURS	0.8432	0.8047	0.7881

5.4 Overall Analysis

We use the area under the curve (AUC) as an evaluation metric to evaluate the performance. The higher the AUC, the better the model performances. The result is shown in Table 2.

From the results we observe that our MKT model achieves the highest performance over three datasets, which verifies the effectiveness of our model. To be specific, our MKT model achieves at least 2% higher results than other baselines on ASSIST09. The main reason why our model performs well is that the input embedding contains more information than the previous model and the ASSIST09 dataset is a typical bipartite graph which nodes interacting each other. As the figure 5 shows, after visualizing the question embedding through t-sne, we find that although there are many questions, we can find that the question embedding with the same skill tag is more intensive in the visualization graph, which reflects the effectiveness of specific meta path extraction information to a certain extent.

So, even in EdNet, our model also gets best result. As for ASSIST12, its question only relates to one skill, our model may not be able to fully capture the relationship between skill and skill, which has also been proved in next ablation study, but our model still performs better than other models.

Compared with PEBG, which also uses the pre-training embedding in the KT task, our model achieves better results with significantly fewer dimension in the embedding process (we only use 10 dimensions while PEBG use 128 dimensions). Moreover, PEBG also adds the similarity between skill and skill, and the similarity between question and question, better results may be further achieved by integrating the similarity calculation in the PEBG into our model.

Compared with AKTHE, which capture the relevance of historical data to the current state by using attention mechanism, our model gets the information between skills and questions by interpretable meta-path and get better performance.

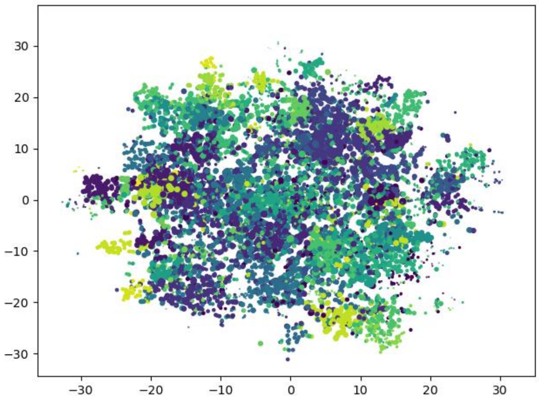


Figure 5. Visualization of Question Embedding (ASSIST09).

5.5 Ablation Study

In this section, we conduct some ablation studies to investigate the effectiveness of our proposed model. We use the single skill embeddings or question embeddings to replace the combination of the skill embeddings and question embeddings. Our experiment shows that the combination of skill and question improves the performance of our model.

As the result shows, the single question embeddings get the best performance in ASSIST12 dataset, but the gap between it and skill-question embedding is very small. After analyzing this dataset, we find that this is related to the characteristics of datasets. The dataset ASSIST12 is different with other two datasets because its question is only related to one skill. Hence in our model, the question embedding can fully contain the information of other questions with the same skill in the meta-path. The MKT-question model does not perform as well as MKT-skill in other datasets where questions generally contain multiple skills, but performs better than MKT-skill in ASSIST12.

Table 3. *The AUC Results over Three Datasets*

	ASSIST09	ASSIST12	EdNet
MKT-S	0.8410	0.7950	0.7815

MKT-Q	0.8260	**0.8055**	0.7784
MKT-S-Q	**0.8432**	0.8047	**0.7881**

6. Conclusion

In this paper, we have proposed a Metapath-based interaction model named MKT, which improves the KT by formulating the question-skill relations as a bipartite graph and introducing Metapath2vec to learn low dimensional embeddings of questions and skills for knowledge tracing. Experiments on three datasets show that MKT significantly improves the performance of existing deep KT models at both skill-level and question-level. Besides, ablation study shows the effectiveness of combination of skill and question for the task of KT, which provides an explanation of its high performance.

Acknowledgements

This work is supported by JSPS KAKENHI Grant Number 18K11597.We would also like to thank four anonymous reviewers for their comments.

References

Chen, H., Perozzi, B., Al-Rfou, R., & Skiena, S. (2018). A tutorial on network embeddings. *arXiv preprint arXiv:1808.02590*.

Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., ... & Heo, J. (2020, July). Ednet: A large-scale hierarchical dataset in education. In *International Conference on Artificial Intelligence in Education* (pp. 69-73). Springer, Cham.

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, *4*(4), 253-278.

Dong, Y., Chawla, N. V., & Swami, A. (2017, August). metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 135-144).

Ghosh, A., Heffernan, N., & Lan, A. S. (2020, August). Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2330-2339).

Grover, A., & Leskovec, J. (2016, August). node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 855-864).

Liu, Q., Huang, Z., Yin, Y., Chen, E., Xiong, H., Su, Y., & Hu, G. (2019). Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, *33*(1), 100-115.

Gan, W., Sun, Y., Peng, X., & Sun, Y. (2020). Modeling learner's dynamic knowledge construction procedure and cognitive item difficulty for knowledge tracing. Applied Intelligence, 50(11), 3894-3912.

Gan, W., Sun, Y., & Sun, Y. (2020, November). Knowledge Interaction Enhanced Knowledge Tracing for Learner Performance Prediction. In *2020 7th International Conference on Behavioural and Social Computing (BESC)* (pp. 1-6). IEEE.

Liu, Y., Yang, Y., Chen, X., Shen, J., Zhang, H., & Yu, Y. (2020). Improving Knowledge Tracing via Pre-training Question Embeddings. *arXiv preprint arXiv:2012.05031*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nakagawa, H., Iwasawa, Y., & Matsuo, Y. (2019, October). Graph-based knowledge tracing: modeling student proficiency using graph neural network. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (pp. 156-163). IEEE.

Perozzi, B., Al-Rfou, R., & Skiena, S. (2014, August). Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 701-710).

Piech, C., Spencer, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. *arXiv preprint arXiv:1506.05908*.

Yang, S., Zhu, M., Hou, J., & Lu, X. (2020). Deep Knowledge Tracing with Convolutions. *arXiv preprint arXiv:2008.01169*.

Yang, Y., Shen, J., Qu, Y., Liu, Y., Wang, K., Zhu, Y., ... & Yu, Y. (2020). GIKT: A Graph-based Interaction Model for Knowledge Tracing. *arXiv preprint arXiv:2009.05991*.

Zhang, J., Shi, X., King, I., & Yeung, D. Y. (2017, April). Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th international conference on World Wide Web* (pp. 765-774).

Zhang, N., Du, Y., Deng, K., Li, L., Shen, J., & Sun, G. (2020, August). Attention-Based Knowledge Tracing with Heterogeneous Information Network Embedding. In *International Conference on Knowledge Science, Engineering and Management* (pp. 95-103). Springer, Cham.

In-process Feedback by Detecting Deadlock based on EEG Data in Exercise of Learning by Problem-posing

Sho YAMAMOTO[a*], Yuto TOBE[a], Yoshimasa TAWATSUJI[b] & Tsukasa HIRASHIMA[c]
[a]*Faculty of Engineering, Kindai University, Japan*
[b]*Global Education Center, Waseda University, Japan*
[c]*Graduate School of Advanced Science and Engineering, Hiroshima University, Japan*
*yamamoto@hiro.kindai.ac.jp

Abstract: Giving feedback to learning activities is one of the most important issues so as to realize adaptive learning. Feedback for the product of the activity (we call it "after-process feedback") has previously been implemented in many interactive and adaptive learning environments. However, feedback during the activity (we call it "in-process feedback") has been hardly implemented. When a learner gets stuck or frustrated during some stage of the process, in-process feedback is much better than after-process feedback. The difficulty in realizing in-process feedback lies in the timing and content of the feedback. To solve this, we developed and implemented affect detection based on EEG data for deciding the timing of the feedback, and knowledge state estimation based on knowledge structure for the content of the feedback. Furthermore, in this study, we realize and evaluate the in-process feedback by detecting deadlocks based on EEG data for learning through problem-posing.

Keywords: Problem-posing, in-process feedback, EEG, knowledge structure, wheel-spinning problem

1. Introduction

We have continued our research to develop and operate a learning environment for problem-posing for arithmetic word problems (Nakano et al., 2000; Yamamoto et al., 2012). The current version of the learning environment is called *Monsakun*, which covers arithmetic word problems that can be solved by one addition and subtraction, those that can be solved by one multiplication and division, and those that use four arithmetic operations (Hirashima et al., 2014). The system diagnoses the problem posed by the learner based on the knowledge structure of the arithmetic word problem. We worked on the practical use of this system in elementary school, and based on the results of this practice, we confirmed that learning by the system is useful for promoting the acquisition of a knowledge structure. We verified that this effect can be obtained not only by students in a regular classroom but also by students in a special needs classroom (Yamamoto & Hirashima, 2016).

Monsakun can be used to diagnose the posed problem based on a model of arithmetic word problems and give feedback to learners. This is feedback on the results of an exercise. We called this type of feedback "after-process feedback." However, this feedback alone is not enough for learners to grasp and modify their errors. This problem is known in intelligent tutoring system (ITS) research as the wheel-spinning problem. In the wheel-spinning problem (1) learners get stuck in the mastery learning loop without any learning occurring, and (2) learners become frustrated when they are repeatedly presented with problems that they obviously cannot solve (Beck & Gong, 2013). Therefore, even in *Monsakun*, to realize useful learning, it is important to realize not only feedback on the posed problem but also feedback on the process of problem-posing (i.e., feedback for the deadlock during learning). We call this feedback "in-process feedback." For this purpose, it is necessary to detect whether the learner is in a wheel-spinning state from two aspects: a stationary point or a cyclic transition of "the knowledge state," and the expression of negative emotions.

We previously performed a model-based analysis of the problem-posing process (Supianto et al, 2017). If the system can generate feedback based on this analysis at the time when the learner's deadlock is detected, it is critical feedback for the learner to resolve the deadlock. Detecting negative emotional states toward an exercise is useful for detecting a deadlock. For detecting (negative) emotional states, attempts have been made to use emotion estimation in learning environments owing to the spread of machine learning and inexpensive measuring devices (Ammar et al, 2010; Alqahtani et al, 2019). For example, Auto Tutor estimates the learner's emotions from the learner's facial expression and uses them as feedback (Graesser et al, 2004; Hussain et al, 2011). However, in terms of the wheel-spinning problem, there is limited research approaching the problem from the perspectives of not only the detection of negative emotions (e.g., Beck & Rodrigo (2014) and Botelho et al. (2019)) but also the detection of the knowledge states of the learner.

Therefore, in this study, to realize in-process feedback, we developed two functions: an EEG-based deadlock detection function using a simple electroencephalograph, and a feedback generating function that points out the cause of the deadlock based on the problem-posing state and the knowledge structure of arithmetic word problem. In Section 2, we described the current version of *Monsakun*, and in Section 3, we describe the design of the in-process feedback. Section 4 introduces the interface of the system used for implementing the in-process feedback. In Section 5, we described simple evaluations and limitations, and in Section 6, we provide some concluding remarks.

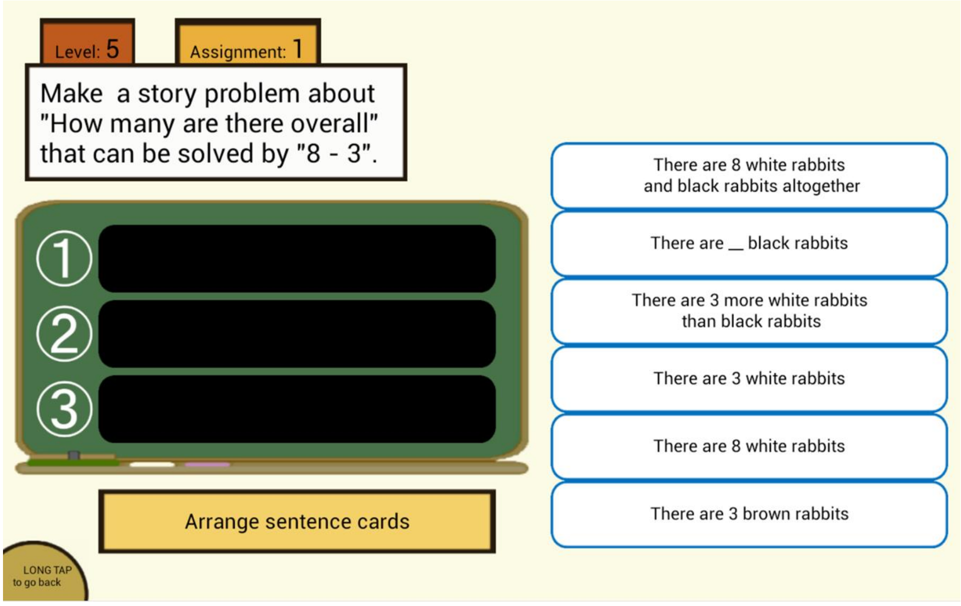
2. Learning Environment based on Knowledge Structure "MONSAKUN"

2.1 Learning by MONSAKUN

A brief description of learning using *Monsakun* is provided in this section. The target domain of this study is a learning environment for problem-posing by arithmetic word problems that can be solved through one addition and subtraction. This system works on tablets.

Figure 1 shows the interface of *Monsakun*, upon on which the learner poses problems. The learner is given a "calculation" and a "story" as constraints when posing a problem, as shown in the upper left part of the figure. On the right side, the learner is given multiple simple sentence cards to pose a problem.

The learner can pose a problem by selecting three correct cards from the given simple sentence cards and arranging them in the proper order. When the learner finishes arranging the cards within the black blank area at the center on the left, the diagnostic button below it becomes active. When the learner taps this button, *Monsakun* diagnoses the problem posed and feeds the results back. This feedback is the after-process feedback. Learners receive this result and deepen their understanding of the knowledge structure by repeating the problem-posing through trial and error.



Make a story problem about "How many are there overall" that can be solved by "8 - 3".

There are 8 white rabbits and black rabbits altogether

There are __ black rabbits

There are 3 more white rabbits than black rabbits

There are 3 white rabbits

There are 8 white rabbits

There are 3 brown rabbits

Figure 1. Interface of *Monsakun* for Problem-posing.

2.2 *Feedback Generation by Knowledge Structure (After-process feedback)*

In this section, we describe the current feedback based on a knowledge structure. Figure 2 shows a knowledge structure of 1-step addition and subtraction arithmetic word problems (Hirashima et al., 2014). Arithmetic word problems that can be solved with a single summation/difference consist of three quantitative concepts. Further, one problem is composed of two independent quantity sentences expressing the existence of the quantitative concept and one relative quantity sentence expressing the relationship between them. Each quantitative concept is expressed based on the quantity (value), what quantity the quantity is (object), and what kind of property it has (predicate). We call a sentence that expresses this single concept of quantity a simple sentence. For example, in the case of "there are two apples," "two" is the quantity, "apple" is the object, and "there is" is the predicate. This simple sentence is an example of an existence sentence as "there is" indicates existence.

There are four story types in a problem that can be solved using 1-step addition or subtraction, i.e., combine, increase, decrease, and comparison. Furthermore, combine and increase are stories that recall addition, and decrease and comparison are stories that recall subtraction. These change the expression of the predicates of the relative quantity concept and combination of the quantity concepts.

In addition, Monsakun has level 1–3 tasks, each of which has a difficulty level set based on the knowledge structure. In level 1, the calculation that recalls by the story required in the assignment and the calculation given in the assignment are same. For example, the learner are required to pose a "Increase" story problem that can be solved by calculating "4 + 5 = ?" This is a story in which the "Increase" story recalls an addition, and the given calculation is also addition. Level 2 has the same condition, although the calculation includes the given value on the left side, such as "4 +? = 9." Level 3 is an assignment in which the calculation of the story and the calculation of the mathematical formula are different. For example, the learner is required to pose a "Increase" story problem that can be solved by calculating "9–5." See Hirashima et al. (2014) for details.

Finally, feedback using this knowledge structure is generated based on whether the problem posed by the learner satisfies the constraints of the above structure. There are a total of five errors that are fed back to the learner. The constraint violations regarding the establishment of the problem are as follows: "object correspondence," "quantity correspondence," and "number of independent quantity sentences and relative quantity sentence." In addition, because "calculation" and "story" are given as assignments in *Monsakun*, there is also a constraint violation of "difference in mathematical formula" and "difference in story." These errors are generated when the learner poses a problem and presses the diagnostic button.

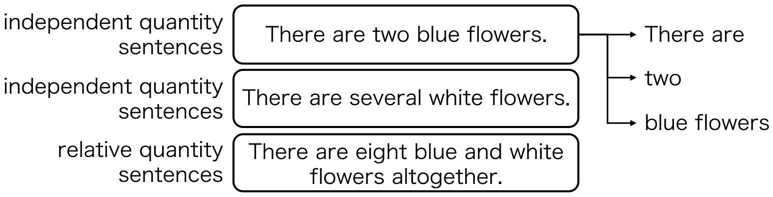

independent quantity sentences
There are two blue flowers.
There are
two
blue flowers
independent quantity sentences
There are several white flowers.
relative quantity sentences
There are eight blue and white flowers altogether.

Figure 2. Triplet Structure of Arithmetic Word Problem by Solving 1-Step Summation/Difference.

2.3 *Research Question*

We have already confirmed that *Monsakun* can provide useful feedback that improves learner understanding. This system only implements after-process feedback based on the diagnosis of the posed problem. However, some learners with a slow learning progress might face difficulty with problem-posing and, thus, be unable to continue to pose the problem. This type of confusion often occurs during the thought process. Therefore, it is crucial to give appropriate feedback at the right time when learners are confused (i.e., negative emotional state), for example, when they face a deadlock.

We previously performed a model-based analysis of the problem-posing process but were unable to generate timely feedback. Process analysis can be performed if even one card is set. Based on this analysis, the assessment can be made every time the learner inserts or removes the card from the blank. However, it is not realistic for the system to provide feedback to the learner at this frequency.

Therefore, we developed in-process feedback, which helps detect negative emotional states based on EEG data and uses its states as deadlocks to generate feedback based on a process analysis.

This is realized using a simple electroencephalograph. After detecting a deadlock, the system generates feedback based on the knowledge structure and the problem-posing situation of the learner. We believe this system can use affect detection to generate feedback that will allow learners to overcome a deadlock.

3. Feedback Design

3.1 Design of Affective Detection

A number of studies on Intelligent Tutoring Systems have considered human emotional states with multifaceted physiological data (Graesser et al, 2004; D'Mello et al, 2007), including an Affective Tutoring System (Ammar et al, 2010). The goal of these studies is to estimate the emotional states of learners based on externally acquired information (e.g., seat pressure and facial expressions from cameras). In recent years, with the advancement of measurement technology, it has become easier to obtain physiological information, and attempts have been made to estimate the emotional aspects of learners using electrocardiography (Alqahtani et al, 2019) and an EEG (Xu et al, 2018). Recent research has attempted to estimate the emotional state of learners from their multifaceted physiological information using a deep learning algorithm (Matsui et al, 2019).

In this section, we describe the development of a model that uses EEGs to detect a deadlock of a learner. The device used was MindWave Mobile 2 manufactured by NeuroSky Inc. The reason for using electroencephalography is that, in general, it allows more freedom of head movement during the measurement than functional magnetic resonance imaging (fMRI) or near-infrared spectroscopy (NIRS) (Miyauchi, 2013). In particular, learners in deadlock may cause head movements (by thinking), and the use of electroencephalography is suitable in this respect. In addition, ordinary electroencephalographs have some issues such as calibration, which imposes a burden on the learner. Wearing the device can affect the learner's affective data. Therefore, in order to minimize these effects, we used a simple electroencephalograph. The acquisition of training data by MindWave Mobile 2 and the model construction by a d-CNN are described below.

3.1.1 Learning Data

We measured the learning data using a simple electroencephalograph when applying *Monsakun* for three university students in the engineering field. Raw data can be acquired from MindWave Mobile 2 as well as values of a low α wave, low β wave, low γ wave, high α wave, high β wave, medium γ wave, θ wave, and Δ wave. We focused on the frequency spectrum, such as alpha and beta waves, assuming that they encode information in the temporal direction as relatively global information, rather than each data in a fine time interval (i.e. sampling rate). The output data included nine emotions: enjoy, hope, pride, anger, anxiety, shame, hopelessness, boredom, and the last one based on the AEQ proposed by Pekrun et al. (2011). The learner answered what kind of feelings they experienced during the exercise from among these nine emotions.

The learner wore a MindWave Mobile 2 device and worked on level 1–3 exercises of *Monsakun* in turn. The learner exercises were recorded on video. Next, the learner answered which of the nine emotions that were felt every 10 s while watching a video of the exercise. In addition, we asked them to answer whether the emotion was caused by the “exercise,” “software UI ,” or “others.” If it was caused by “software UI,” and “others,” it was deleted from the training data.

In addition, the emotional response also converted the positive emotions enjoy, hope, and pride into “a state in which the exercise is proceeding smoothly.” We also transformed the negative emotions anger, anxiety, shame, hopelessness, and boredom into “a state in which the exercise is not proceeding smoothly (deadlock)” Therefore, the actual output values are 1 for “the state in which the exercise is proceeding smoothly” and 0 for “deadlock” This is listed as "n/p" in the table. Therefore, the learning data are as shown in Table 1. The number of data are 572 for level 1, 557 for level 2, and 1079 for level 3.

Table 1. *Example of Training Data for Affective Estimation on Monsakun*

Delta	High Alpha	High Beta	Low Alpha	Low Beta	Low Gamma	Mid Gamma	Theta	n/p
207,877	6,426	7,024	12,333	14,276	1,040	1,094	56,468	0
74,278	15,553	2,553	8,419	7,101	3,227	580	18,357	1

3.1.2 Model Generation by Deep Learning (3-CNN)

In this section, we describe the construction of a model for deadlock estimation based on an EEG, created using the learning data described in the previous section. Deep learning was used as the machine learning because it was assumed that the activation of human emotions used is closely related to the movement shown in the EEG. The learner of machine learning was set to 3 hidden layers and 10 nodes for each layer. In addition, the dropout rate was set to 20% to prevent overfitting. Next, the activation function was set to tanh because the activation of human emotions is gradual, and the loss and evaluation functions were set to the mean square error. These settings are experimental. The batch size is a standard value of 32, and the data were divided into 95% training data and 5% test data because few learning data were prepared this time.

Next, the data examined for the optimal model construction are described. In this study, we verified the model accuracy by changing the gradient method, number of epochs, and learning rate. Seven gradient methods were examined: SGD, Adadelta, Adagrad, Adam, Adamax, RMSprop, and Nadam. When the epoch was examined experimentally using each gradient method, an overfitting occurred at 1000 epochs or more; thus, we decided to examine four numbers of epochs: 100, 300, 500, and 800. The learning rate was set to 0.1, 0.05, and 0.01. These parameters were experimental.

The procedure used for building the model is as follows: After using the above learner, the learning rate was first examined by fixing the number of epochs to 300 using each gradient method. Next, using the most accurate learning rate, we examined the model with 100–800 epochs. At this time, the epoch with the highest accuracy in each gradient method was used as the representative value. Finally, the epoch with the highest accuracy among each gradient method was adopted.

The above operations were carried out at each level of 1–3, and deadlock detection model for each level were created. Table 2 shows the data adopted for each level of 1–3.

Table 2. *Results of Machine Learning in Levels 1–3*

	Learning Rate	Epoch	Gradient Method	Accuracy	Loss
Level 1	0.1	500	Adadelta	0.778	0.134
Level 2	0.05	100	RMSprop	0.778	0.183
Level 3	0.01	300	Nadam	0.704	0.152

3.2 Design of Feedback based on Posing Problem and Knowledge Structure

The design of feedback during the exercise using the knowledge structure is as follows. The system generates feedback when a learning deadlock is detected using the deadlock detection model based on the EEGs and machine learning described in the previous section. Therefore, the system should assess the in-process problem, rather than the after-process problem. It is based on the model-based analysis of the problem-posing process that has already been implemented (Supianto et al, 2017).

Monsakun can detect a constraint violation if some cards are answered by the learners. Table 3 shows the correspondence between this constraint violation and feedback. Such feedback is only generated when the EEG diagnosis detects a deadlock and when there are fewer than three cards answered. First, the system checks the number of cards answered by the learner when the EEG detects a deadlock. The type of simple sentence card answered is detected. The system then confirms the story given in the assignment. The sentence of the feedback may change depending on this given story. Finally, whether the answered card satisfies the conditions for a feedback generation is detected and the feedback shown in the feedback sentence of Table 3 is generated. As an example, suppose there is one answered card and it is a relative quantity sentence. If the answered card is difference from given story by assignment, the learner will receive feedback that says, "Be careful about the type of story."

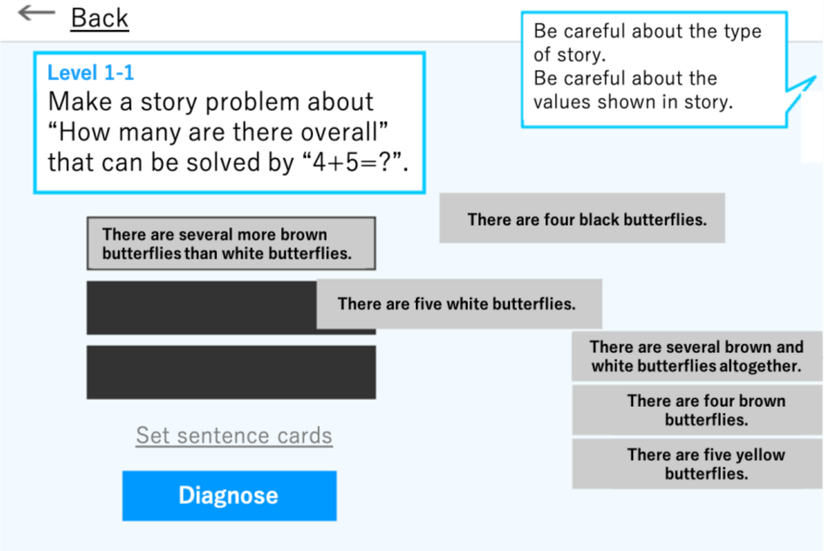
Table 3. *Correspondence between Type of Answered Cards, Constraint Violations and Feedback Sentences*

Number of answered cards	Kind of Card	Given story	Condition	Feedback Sentence
1	Relative quantity sentence	Irrelevant	Difference of given story	Be careful about the type of story.
			Same as given story	You're doing good.
	Independent quantity sentence	Irrelevant	Set cards not used for assignments	Be careful about the objects shown in story.
			Set cards used for assignments	You're doing good.
	Irrelevant	Irrelevant	Set cards not contained correct value	Be careful about the values shown in story.
			Set cards contained correct value	You're doing good.
2	Two independent quantity sentences	Combine and Difference	Each cards' objects are same	Be careful about the objects shown in story.
			Each cards' objects are different	You're doing good.
		Increase and Decrease	Each cards' objects are same	You're doing good.
			Each cards' objects are different	Be careful about the objects shown in story.
	Relative and independent quantity sentences	Irrelevant	Relative quantity sentences is not correct card	Be careful about the type of story.
			Objects of relative and independent quantity sentences are different	Be careful about the objects shown in story.
			n/a	You're doing good.
	Irrelevant	Irrelevant	Set cards contained incorrect value	Be careful about the values shown in story.
			Set cards contained correct value	You're doing good.

4. MONSAKUN Affective

An outline of the system for implementing the in-process feedback is described in this Section. We call this system *Monsakun Affective*. As the basic operation, the level of assignment and the assignments implemented are the same as in the conventional system introduced in Section 2. However, the learner must wear an electroencephalograph before starting the exercise for receiving new feedback. The function used to detect a deadlock is being developed as a Web API. Therefore, *Monsakun Affective* sends the data from the electroencephalograph to the Web API every second, and the Web API returns the result of the deadlock judgment based on the model in Section 3.1.

Here, we describe the exercise procedure of *Monsakun Affective*. The learner wears an electroencephalograph, confirms that the data can be measured without any problems, and then logs into the system. The level selection interface is then displayed, and the learner selects the level to work on from among levels 1–3. The system then displays the practice interface shown in Figure 3. The procedure of the exercise and feedback after the problem-posing is the same as in a usual *Monsakun,* although the feedback based on an EEG is displayed in the upper-right. If the learner feels confused, the balloon in the upper right will display the feedback described in Section 3.2.



Level 1-1
Make a story problem about "How many are there overall" that can be solved by "4+5=?".

Be careful about the type of story.
Be careful about the values shown in story.

There are several more brown butterflies than white butterflies.

There are four black butterflies.

There are five white butterflies.

There are several brown and white butterflies altogether.

There are four brown butterflies.

There are five yellow butterflies.

Figure 3. Interface of *Monsakun Affective* for In-Process Feedback.

5. Experimental Use

5.1 Procedure

The purpose of this experiment is to confirm whether in-process feedback is properly generated using the developed prototype system. The subjects were five engineering college students, who differed from the students who acquired the learning data.

First, the subject was instructed on how to use *Monsakun Affective* and the experiment procedure. Next, the subject used the developed prototype system to work on an assignment of levels 1–3. At this time, the exercise was recorded on video. Next, the subject was asked to answer whether the timing and content of the feedback implemented were appropriate while watching the video. There were four answers regarding the timing: "appropriate," "early," "slow," and "not necessary." There are two types of content answers, "appropriate" and "inappropriate." Finally, the subject answered the usage questionnaire.

5.2 Evaluation of Timing and Contents

Table 4 shows various exercise logs of the system. All values are averages for all subjects by level. If we check the exercise time, level 3 is overwhelmingly large. The number of feedbacks is divided into after-process feedback and in-process feedback. The number of all feedbacks are only for mistakes and do not include the number of feedbacks for correct answers. There are more the number of feedback for level 3 than each number of feedback for level 1 and 2 as well. The number of assignments for each level 1-3 is 10 questions. Therefore, the number of posed problems for Levels 1 and 2 is almost the same as the number of assignments. However, the number of posed problems for level 3 is about three times the number of assignments. The number of steps is the number of times the card is put in and out of the blank. It takes a minimum of 3 steps to pose the correct problem. Therefore, the minimum number of steps for each level is 30. More than this number of steps is the steps in which the learner performed various thinking activities other than giving the correct answer, and this is described as the number of search steps. The number of steps was lowest at level 2 and highest at level 3.

Moreover, there were more number of in-process feedbacks than the number of after-process feedbacks. *Monsakun* assessed that many feedbacks are required at the in-process timing in addition to the after-process feedback.

Table 5 provides answers regarding the adequacy of the in-process feedback timing and content. The subjects judged that the timing of in-process feedback was appropriate for about 70%.

Although there is a possibility of an overfitting with machine learning, the possibility of an overfitting is low because the subject is not a student who acquired the learning data and logs of multiple learners were combined into such data. Although 10% to 20% of the feedback was deemed “unnecessary,” this level decreased to less than 10% as the difficulty increased. We suspect that this result is because the learner was received feedback when he/she was thinking in the same way as the content of the feedback. Currently, the system generates feedback the moment it detects a deadlock; however, it may be better to generate feedback later.

Next, we consider the answers to the feedback content. Approximately 90% of the feedback content was considered appropriate. The feedback that was deemed the most inappropriate was “think about the values and objects shown in the story.” It is possible that the subject did not consider it an error because they merely overlooked objects and values. It was shown that in-process feedback not only determined that *Monsakun* needed it, but that the learner also determined it to be meaningful.

Table 4. *Logs of Exercise using Monsakun Affective (N=5)*

	Exercise time	Number of feedbacks		Number of posed problems	Number of steps	
		After-process	In-process		Total	Exploring Steps
Level 1	4m48s	1.4	15.4	11.4	55.8	25.8
Level 2	4m43s	1.0	16.8	11.0	44.6	14.6
Level 3	16m28s	17.4	40.8	27.4	156.6	126.6

Table 5. *Result of Timing and Content of In-process Feedback in Levels 1–3*

	Timing				Content	
	appropriate	early	slow	not necessary	appropriate	Inappropriate
Level 1	0.66	0.14	0.01	0.18	0.87	0.13
Level 2	0.70	0.19	0.00	0.11	0.87	0.13
Level 3	0.77	0.15	0.01	0.07	0.91	0.09

5.3 *Questionnaire*

The contents and results of the questionnaire are shown in Figure 4. There are six answers: “strongly agree,” “agree,” “somewhat agree,” “somewhat disagree,” “disagree,” and “strongly disagree.” However, in Q2 to Q4, there are 6 answers: "Very difficult", "Difficult", "Somewhat difficult", "Somewhat easy", "Easy", and "Very easy". In the Figure 4, the graph is drawn by replacing "very difficult" with "strongly agree." First, all learners were able to concentrate on the exercises, and the simple electroencephalograph was not in the way. Most of the respondents answered that levels 1 and 2 were easy, but level 3 was difficult, even for college students. In addition, as an important result, many of the subjects pointed out that it was difficult to notice new feedback. This led to the negative answers to Q6 and 7.

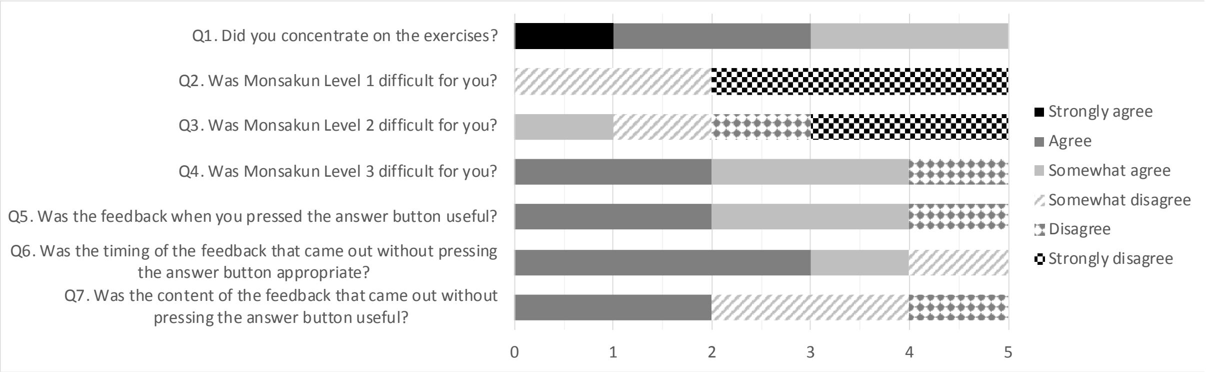


Figure 4. Contents and Results of Questionnaire (N = 5).

5.4 Discussion and Limitation

We were able to develop a system that can detect a deadlock in a learner and return feedback by using the knowledge structure and affective detection. Detecting a deadlock based on EEG data using a simple electroencephalograph is considered to be sufficiently practical, with about 70% of the results showing an appropriate timing. Overfitting may have occurred as a result of machine learning. However, the learning data used is a combination of data from multiple learners, and the test subjects were different from the learners who collected the learning data. Therefore, we believe that the possibility of overfitting is low. Of course, this model should be verified in the future. In addition, we considered a simplification of the model by making the objective variable binary also contribute to the system. The objective variable could be simplified because the system has a sufficient knowledge structure and can estimate the error of the learner during an exercise. Thus, even if the system cannot detect detailed emotions, meaningful feedback can be generated that will clear the deadlock of the learner.

From the exercise log, it was found that the subjects were conducting problem-posing activities for arithmetic word problems through trial and error. Especially level 3 is remarkable. In *Monsakun,* the number of steps greater than the shortest number of steps indicates a process of exploratory thinking rather than answering the correct answer. Therefore, all these steps can be the target of feedback. It is possible to give feedback with all these steps, but it hinders the exercise and is not realistic. In-process feedback reduces this feedback to around 40% on average. In contrast, the number of after-process feedback is too small. Considering that about 70% of in-process feedback was effective, it is considered that sufficiently useful feedback was realized. Based on the above, we confirmed the feasibility of in-process feedback. This research is aimed at arithmetic word problems; however, if the system has a significant knowledge structure, the same effect can be expected.

However, the interface of the system is inappropriate. It is necessary to consider how to present new feedback. It is also difficult to verify that the introspection report at the time of acquisition of the learning data was correct (D'Mello et al, 2007) because in the case of this study, it takes time from the end of the exercise to the introspection report to be created. Learners must recall their feelings during the exercise. Moreover, it is also necessary to understand that learners occasionally do not report honestly about their affective states. Furthermore, in this study, we experimentally adjusted the hyperparameters of machine learning. By contrast, Young et al. (2012) proposed a method using a genetic algorithm for adjusting the hyperparameters of a convolutional neural network consisting of three layers. We are going to plan that this method will be used to determine the hyperparameters in the future.

6. Concluding Remarks and Future Work

We worked on the realization and verification of in-process feedback to resolve deadlocks in learning by problem-posing. This feedback function detects a deadlock in a learner based on emotion detection using brain wave data and machine learning and identifies the cause of the deadlock based on the knowledge structure and state of the problem-posing. The existing system realizes the assessment and feedback of a posed problem based on the knowledge structure. This is feedback on the posed problem and can be said to be a post-process feedback, which is useful for learners to modify errors.

However, when using such a system, if the learner becomes confused while thinking about an exercise, this feedback will not work properly. Therefore, we aimed to develop in-process feedback for learners who are find themselves in a deadlock during an exercise. Previously, we were able to analysis the problem-posing process, and thus generating feedback at the right time is a significant challenge. We used EEG-based affect detection to solve this challenge. The accuracy of the developed deadlock detection system was approximately 70%, and the appropriateness of the feedback sentence was approximately 90%. This result shows that in-process feedback can be realized effectively in resolving the deadlock of a learner. We believe this will provide insight into solving the wheel-spinning problem in an ITS.

In contrast, we must think about how to provide such feedback. A learner pointed out that it was difficult to notice the suggested feedback during the exercise. Furthermore, developing machine learning models should also be considered. For example, it is necessary to consider the accuracy of the introspection report of the learner when acquiring the learning data. We also plan to improve the value of hyperparameters. As additional future work, we plan to verify (1) the difference of learning gain between in-process and after-process feedback, and (2) the difference in the effect of each feedback when EEG is replaced with another device.

References

Alqahtani, F., Katsigiannis, S., & Ramzan, N. (2019, August). ECG-based affective computing for difficulty level prediction in intelligent tutoring systems. In 2019 *UK/China Emerging Technologies (UCET)* (UK), Glasgow (pp. 1-4). IEEE.

Ammar, M. B., Neji, M., Alimi, A. M., & Gouardères, G. (2010). The affective tutoring system. *Expert Systems with Applications*, 37(4), 3013-3023.

Beck, J. E., & Gong, Y. (2013, July). Wheel-spinning: Students who fail to master a skill. *In International conference on artificial intelligence in education* (USA), Memphis, Tennessee (pp. 431-440). Springer, Berlin, Heidelberg.

Beck, J., & Rodrigo, M. M. T. (2014, June). Understanding wheel spinning in the context of affective factors. In *International conference on intelligent tutoring systems* (USA), Honolulu, HI (pp. 162-167). Springer, Cham.

Botelho, A. F., Varatharaj, A., Patikorn, T., Doherty, D., Adjei, S. A., & Beck, J. E. (2019). Developing early detectors of student attrition and wheel spinning using deep learning. *IEEE Transactions on Learning Technologies*, 12(2), 158-170.

D'Mello, S., Picard, R. W., & Graesser, A. (2007). Toward an affect-sensitive AutoTutor. *IEEE Intelligent Systems*, 22(4), 53-61.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. Behavior Research Methods, *Instruments, & Computers*, 36(2), 180-192.

Hirashima, T., Yamamoto, S., & Hayashi, Y. (2014, June). Triplet structure model of arithmetical word problems for learning by problem-posing. In *International Conference on Human Interface and the Management of Information* (Greece), Heraklion, Crete (pp. 42-50). Springer, Cham.

Hussain, M. S., AlZoubi, O., Calvo, R. A., & D'Mello, S. K. (2011, June). Affect detection from multichannel physiology during learning sessions with AutoTutor. In *International Conference on Artificial Intelligence in Education* (New Zealand), Auckland (pp. 131-138). Springer, Berlin, Heidelberg.

Matsui, T., Tawatsuji, Y., Fang, S., & Uno, T. (2019, June). Conceptualization of IMS that estimates learners' mental states from learners' physiological information using deep neural network algorithm. In *International Conference on Intelligent Tutoring Systems* (Jamaica), Kingston (pp. 63-71). Springer, Cham.

Miyauchi, S. (2013). Non-invasive study of human brain function and psychophysiology (2nd edition), *Japanese Psychological Review*, 56(3), 414-454. (in Japanese)

Nakano, A., Hirashima, T., & Takeuchi, A. (2000). A Learning environment for problem posing in simple arithmetical word problem, In *Proceedings of the 16th International Conference on Computers in Education*,（pp. 91-98).

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary educational psychology*, 36(1), 36-48.

Supianto, A. A., Hayashi, Y., & Hirashima, T. (2017). Model-based analysis of thinking in problem posing as sentence integration focused on violation of the constraints. *Research and Practice in Technology Enhanced Learning*, 12(1), 1-21.

Xu, T., Zhou, Y., Wang, Z., & Peng, Y. (2018). Learning emotions EEG-based recognition and brain activity: A survey study on BCI for intelligent tutoring system. *Procedia computer science*, 130, 376-382.

Yamamoto, S., & Hirashima, T. (2016). A case study of interactive environment for learning by problem-posing in special classroom at junior high school. In *Proceedings of the 24th International Conference on Computers in Education* (India), Mumbai (pp. 282-287).

Yamamoto, S., Kanbe, T., Yoshida, Y., Maeda, K., & Hirashima, T. (2012, November). A case study of learning by problem-posing in introductory phase of arithmetic word problems. In *Proceedings of the 20th International Conference on Computers in Education* (Singapore), (pp. 25-32).

Young, S. R., Rose, D. C., Karnowski, T. P., Lim, S. H., & Patton, R. M. (2015, November). Optimizing deep learning hyper-parameters through an evolutionary algorithm. In *Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments* (USA), Austin, Texas (pp. 1-5). Association for Computing Machinery.

Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

An Improved Model to Predict Student Performance using Teacher Observation Reports

Menna FATEEN*, Kyouhei UENO & Tsunenori MINE
Kyushu University, Fukuoka, Japan
*menna.fateen@m.ait.kyushu-u.ac.jp

Abstract: Predicting students' performance is a highly discussed problem in educational data mining. A tool that can accurately give such predictions would serve as a valuable resource to teachers, students, and all educational stakeholders as it would provide essential insights. Students can be further guided and fostered to achieve their optimal learning goals. In this paper, we propose an improved method to predict students' performance in entrance examinations using comments that their cram school teachers took throughout lessons. Teachers in these cram schools observe their students' behavior closely and give reports on the efforts taken in their subject material. We compare our previous model with a new and improved one to show that teachers' comments are qualified to construct a reliable tool capable of predicting students' grades efficiently. These methods are new since studies previously focused on predicting grades mainly using student data such as their reflection comments or earlier scores. Our improved experimental results show that using this readily available feedback from teachers can predict students' letter grades with an accuracy of 68%.

Keywords: Text mining, student grade prediction, teacher observation reports, machine learning

1. Introduction

Grade prediction is one of the most prominent challenges in the educational data mining community with a wide range of diverse solutions (Bretana et al., 2020; Sweeney et al., 2015). A tool that can accurately predict students' future grades is considered a powerful means that can provide valuable and beneficial insights to all educational stakeholders. These insights include early identification of at-risk students, factors that affect student performance and more. Grade prediction tools usually adopt either historical student data such as their previous grades or require that students write comments after class (Luo et al., 2015).

On the other hand, in educational systems, a trend can also be observed where educational programs evolve towards more active learning. Teachers spend more time observing their students and designing class material that encourage engagement. In innovative universities such as the Minerva Schools at KGI in the United States, the student/teacher ratio is low, and students receive written feedback from their teachers daily that clarifies any confusion, reinforces strong points, and gives more specific advice and guidance (Kerrey, 2018; Han & Xu, 2020). This trend produces a large amount of unstructured data generated by the teachers in form of reports and comments.

In this paper, we introduce an improved grade prediction model that can adequately exploit unstructured data to provide essential insights to students and teachers alike. We improve upon our previous model (Fateen & Mine, 2021) that employs teacher reports provided to us by a cram school in Japan. Cram schools are specialized in providing extra and more attentive education for students who want to achieve certain goals, particularly studying for high school or university entrance exams. The cram school sector in Japan, or 'juku', is a very large and influential one, bringing in billions of yen in profit each year. Research in the area of cram school education is needed and encouraged (Lowe, 2015).

Our two main developed grade prediction models utilize both classic and state-of-the-art techniques in Natural Language Processing (NLP) to capture the meanings of the teachers' comments. We use these vectorized reports as explanatory variables for our machine learning regression models. The experimental results of our improved model proposed in this paper show that when adding teachers' reports to the regular student exam scores, we can correctly predict their letter grade with an accuracy up to 68%. To sum up, our contributions can be outlined as:

- We propose an improved grade prediction method and compare it with our previous method. Both models use teacher observation reports represented using classic (bag of words, term-frequency inverse-document-frequency) and state-of-the-art (BERT) methods.
- We conduct extensive experiments on real data sets so as to prove the capability of teachers' reports for building accurate grade prediction models.

Finally, to the best of our knowledge, this is the first study to mine teacher observation reports to predict student grades. Our research and experimental results demonstrate the potential that these teacher observation comments have in predicting students' total scores and final letter grades.

2. Related Work

Data mining techniques are being increasingly adapted in many different fields from engineering to healthcare (Chen et al., 2017; Rushdi et al., 2020). Needless to say, many diverse solutions to significant problems in the educational community have been introduced that utilize machine learning and artificial intelligence. Educational data mining solutions vary widely from course recommendation systems (Ma et al., 2017) to automatic feedback models (Makhlouf & Mine, 2020). More specifically, many studies have been dedicated to build precise grade prediction models using various techniques such as predicting next term grades using cumulative knowledge bases (Morsy & Karypis, 2017). Bydžovská (2016) used two approaches to predict students' final grades. The first approach utilized students' social behavior while the second approach used a collaborative filtering technique where the final grade was predicted based on previous achievements of similar students. Both approaches had similar average results and the paper described each approach's advantages and disadvantages.

Several methods that utilize natural language processing have been developed to predict student performance especially since the field of NLP has seen many significant breakthroughs over the past few years. These NLP methods have been proven to have the capacity to contribute to accurately predicting students' success or failure over the information usually obtained from fixed-response items (Robinson et al., 2016). Luo et al. (2015) proposed a method that predicts students' grades based on free-style comments taken according to the PCN categorical method (Goda & Mine, 2011). Word2Vec embeddings were adopted to reflect the meanings of the students' comments. This was followed by an artificial neural network to predict the student grades. Their results showed a correct rate of 80%.

Teacher comments have been used by Jayaraman (2020) not to predict students' grade, but to detect those students who are at risk of dropping out of university. Sentiment analysis was used in their study to first extract the positive and negative words from the advisors' notes and then fed into a model as features. Their method achieved a 73% accuracy at predicting dropout.

3. Data Description

The data used in our experiments were obtained from a cram school in Fukuoka, Japan. No student names or any other identifiers were included to ensure confidentiality. We obtained the teachers' reports monthly as CSV files in addition to the students' scores of their exams taken at their regular schools. The final dataset compromised 11,960 reports for 167 students over the period from May to October 2020.

3.1 Monthly Reports

In addition to the teachers' comments, each report also contained the class date, subject code, understanding, attitude and homework scores. The feature sets adopted in our experiments are discussed in detail in Section 4.2. The main explanatory variable, however, is the teachers' comments. The average length of these comments was 96 characters and a word cloud of the most used words in the teacher's comments is shown in Figure 1.

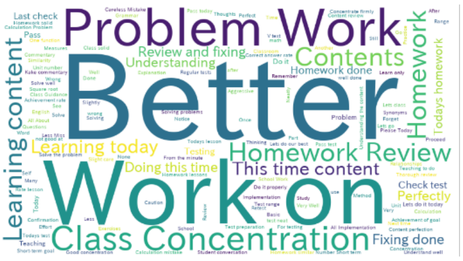


Figure 1. Word Cloud of Translated Comment Words.

The comments are originally written in Japanese and the words were translated using Google Translate. From the cloud, it can be perceived that teachers tend to encourage and energize their students by using words such as “better” and “work on.” Moreover, the words used in the comments depend on the context or class subject to some degree. For example, the expression “calculation problem” is likely to be used frequently in math lessons.

Students in the cram school take different lessons for each subject. The lessons fall under the five main subjects: Japanese, Mathematics, Science, Social Studies and English. Since the main goal of our model is to predict a student’s total score, reports in all the five subjects were required. In our first model we thus eliminate from our experiments the students that did not attend lessons for all the subjects. However, in the second model, we also conduct experiments with a compensation model to try to estimate the missing subject scores. This will be discussed in Section 5.3.

Table 1 shows the number of reports that fall under each subject. The number of total reports for each student varied depending on the number of classes attended. The average number of total reports recorded for each student was 82 reports with a maximum of 206 and a minimum of 24 reports. A histogram that visualizes the distribution of the number of reports for students is shown in Figure 2.

Table 1. *The Number of Reports in Each Subject along with the Standard Deviation and Mean of the Students’ Scores in Each Subject*

	Japanese	Math	Science	Social Studies	English
Number of Reports	1157 (9.7%)	3547 (29.6%)	2428 (20.3%)	1669 (14%)	3159 (26.4%)
Stand. Dev.	11.85	16.62	20.33	16.93	18.47
Mean	54.56	49.41	48.10	44.74	47.12

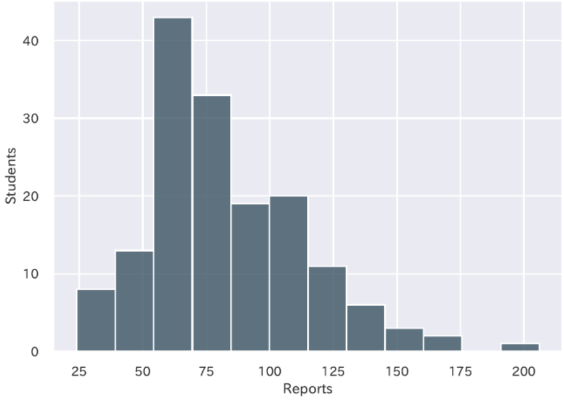


Figure 2. Histogram of the Number of Reports Distributed among Students.

3.2 *Test Scores*

Students attending the cram school are normally registered in many different schools. The results of their examinations regularly taken at school were also provided to us. These scores were what we considered as student data and would be traditionally used as the main feature to predict their performance in the entrance exam.

In our proposed models, we adopted a supervised machine learning approach where training data needs to be labeled with the required outputs for each input. This is required so that the model can alter its learning function based on the correct results. Since the students' actual performance in their entrance examinations is unattainable to us, we used the students' results in their simulation exams as the model labels. Histograms for each subject scores were plotted to visualize the distribution as shown in Figure 3. We can observe that the distributions are approximately bell-shaped and seem symmetric about the mean so we can assume that they follow a normal distribution. To show how dispersed the values are, we display the standard deviation for subject scores in Table 1.

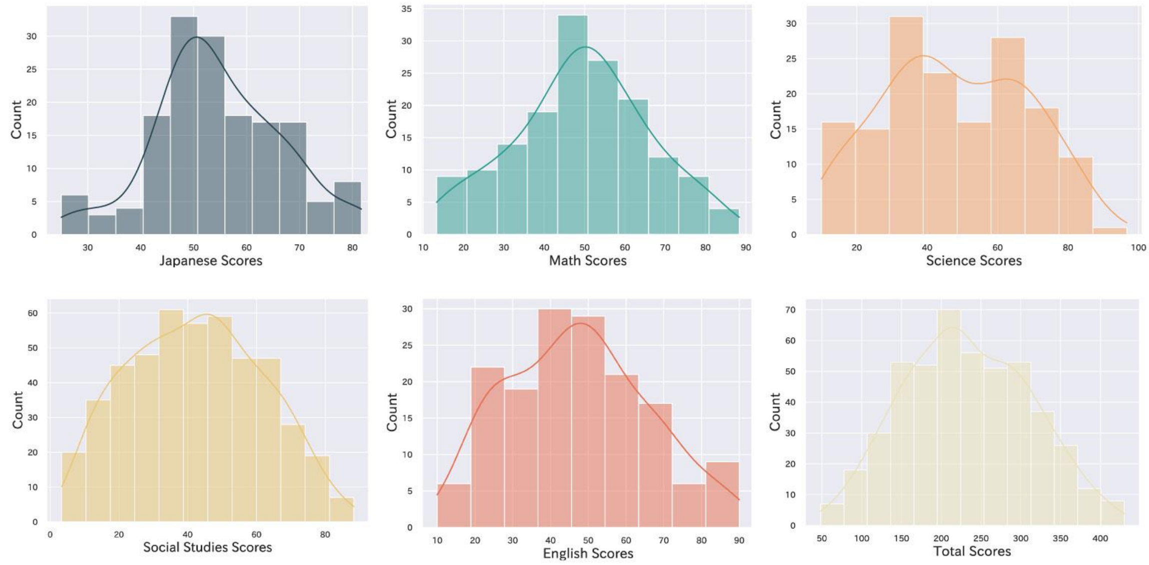


Figure 3. Distribution of Simulation Test Scores.

4. Methodology

4.1 Proposed Models: Reports Model and Students Model

Two main approaches were taken in our experiments to predict students' final grades. In our previous model called "Reports Model," each report is treated as an independent instance. For a specific subject *s* ∈ S, where S = {Japanese, Math, Science, Social Studies, English}, a student *i* can attend a variable number *t* of lessons and therefore have *t* reports. To predict the subject score of student *i*, each report *t* is fed into a regression model separately and an ordered list $X_{i,s,t}$ of predicted scores for student *i* is obtained. Finally, to determine the estimated predicted score of a subject, we use $SubjectScore_{pred,i,s} = Median(X_{i,s,t})$. The total predicted score ($TotalScore_{pred,i}$) can then be estimated by $TotalScore_{pred,i} = \sum SubjectScore_{pred,i,s}$ for s ∈ S. These steps are illustrated in detail in Figure 4.

We introduce another model called "Students Model" as an improved one in this paper. In the model, a separate regression model for each subject judges a student's performance based on all their reports combined in the specified subject. In this model, we aim to encapsulate a student's performance in one vector since the performance can naturally vary from class to class due to different factors. The elements in each vector of the Students Model depend on the feature set used, which is discussed in the next section.

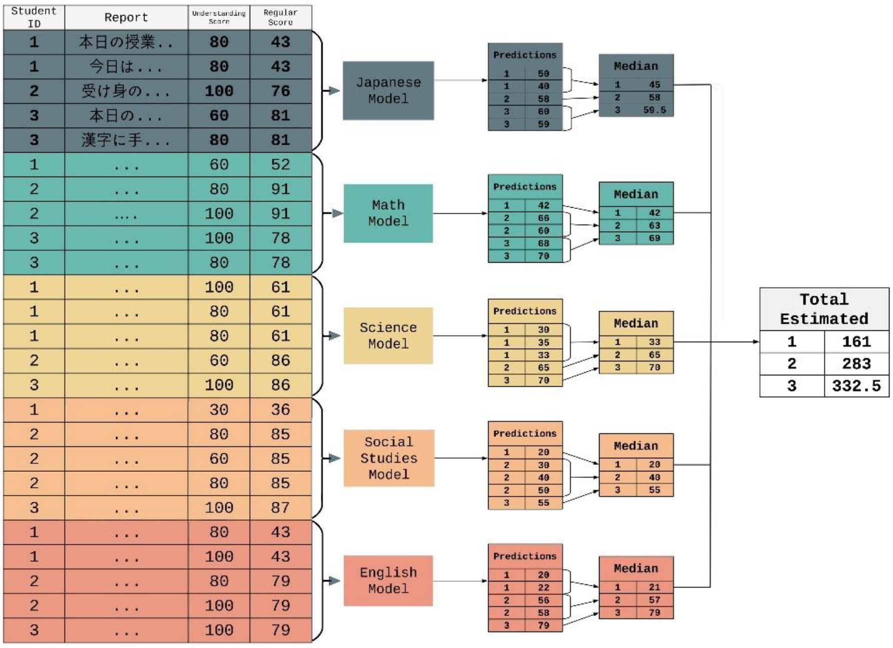


Figure 4. Reports Model Architecture. Each subject model takes a report instance which as input which includes the teachers' comments, an assessment of their understanding given by a score of either (0-30-60-80-100) and an attitude score of (1-2-3-4).

4.2 Feature Selection

For the sake of comparison, we adopt three primary feature sets in our experimental settings. The first feature set, FS_1 consists of teachers' report contents only as the main explanatory variables. After each lesson, the teacher provides a report for each student consisting of written comments based on their observations, an assessment of their understanding given by a score of either (0-30-60-80-100) and an attitude score of (1-2-3-4). In the Reports Model, FS_1 used the complete report as attributes except for the homework score since the score was not included in more than 36% of the reports. Conversely, since the Students Model aims to encapsulate the meaning of all provided reports, a more statistical approach had to be taken. Therefore, in the Students Model, FS_1 consists of the number of classes taken by the student, the first, second and third quartile of the understanding score and finally the teachers' comments.

The second feature set, FS_2, consists of student-related data only. In the Reports Model, this set consisted of their gender and the score of their regularly scheduled exam at school, which we call the students' regular scores or the regular score. Since we predict each subject score separately, the regular score corresponds to the subject score. The gender attribute was omitted for the Students Model since the Pearson correlation coefficient between the gender and the regular score is 0.12, which shows no statistical significance while the correlation coefficient between the regular score and the simulation score is 0.80. Moreover, in the Students Model, all 413 students' regular scores were taken into consideration since this feature set does not depend on subject reports or classes. Finally, we investigate using both teachers' reports and the regular scores to verify whether adding teachers' reports contributes to the accuracy of the prediction model or not. FS_3 therefore is a concatenation between each model's FS_1 and $FS_{2.}$ Figure 5 shows the vector elements when using FS_3 in the Students Model.

4.3 Natural Language Processing

Various ways exist to represent texts to convey the original meanings and prevent information loss. We chose to represent the teachers' comments using three main NLP techniques. In the Reports Model, a TFIDF (term frequency-inverse document frequency) vectorization method is adopted and compared with BERT embeddings. In the Students Model, we use the classic bag of words method along with BERT.

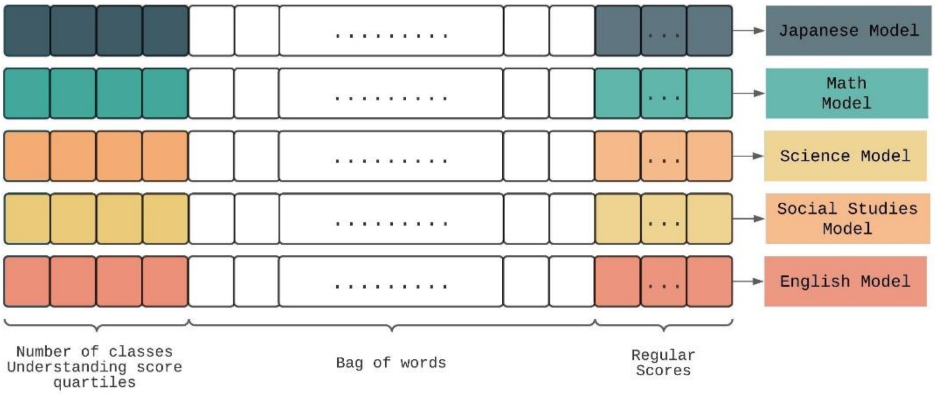


Figure 5. Vectors using FS$_3$ in the Students Model. Each vector in this case consists of (the number of classes taken by the student, the first, second and third quartile of the understanding score, the teachers' comments and the students' regular scores)

4.3.1 Bag of Words

The first essential step in transforming a text into a numerical representation is preprocessing the text. For English sentences, this step begins with splitting the sentences into words or tokenization. Tokenization in languages such as English can be simply done by splitting the sentence strings at each space. However, for Japanese, this step is merged with the next, which is morphological analysis, since there are no spaces in Japanese sentences. We use Mecab (Kudo, 2006), a Japanese tokenizer and morphological analysis tool to extract the nouns and verbs from the teachers' comments. We then use these extracted words to build the corpus of vocabulary for Bag of Words (BOW). In a BOW vector, each element corresponds to a word in the corpus and represents its frequency in the sentence. In the Students Model, we employed sklearn's CountVectorizer to build the BOW vector (Pedregosa & Karypis, 2011). The length of the BOW vector is 6033 words. In the Reports Model, we experimented with a variant of BOW, TFIDF, where high weights are given for those terms that occur often in a particular sentence or document but rarely in other documents.

4.3.2 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a novel technique of pre-training language representations presented by Google (Devlin et al., 2018). On many NLP tasks, BERT obtains state-of-the-art results. A pre-trained BERT model is essentially a general-purpose language understanding model trained on a large corpus like Wikipedia. This pretrained model can then be utilized for downstream NLP tasks. When building both the Reports and Students Models, we used the BERT model pretrained by Inui Laboratory at Tohoku University (cl-tohoku, 2019). The model was trained with the same configuration as the original BERT and the pretraining corpus used was Japanese Wikipedia. We used the [CLS] token embeddings as our BERT embeddings to transform the teachers' comments into constant length vectors.

4.4 Evaluation Metrics

To evaluate our experiments, we adopt two main evaluation metrics: MAE (Mean Absolute Error) and PTA (Percentage by Tick Accuracy). MAE is calculated using the following formula:

$$MAE = \frac{1}{n}\sum |score_{pred,i} - score_{true,i}|$$

, where $score_{true,i}$ is the actual score that student i obtained. In the Reports Model, $score_{pred,i}$ for subjects is calculated by taking the median, as explained in Section 4.1, while in the Students Model the subject score is predicted directly by the regression model. The total score is the summation of all subject scores predicted by either model.

PTA has been used in previous studies to evaluate grade prediction models. Since students receive letter grades for their total score, we map the estimated total score to its corresponding letter grade according to the percentages as shown in Table 2. A tick, as specified by (Polyzou & Karypis, 2016) is the difference between two successive letter grades. In the experiments we employ PTA_0 and PTA_1, which mean the model successfully predicted the exact letter grade or predicted it incorrectly but with 1 tick away from the true grade (e.g., B vs C), respectively.

Table 2. *Letter Grades and their Corresponding Percentages*

Letter Grade	S	A	B	C	D	F
Percentage	90-100%	80-89%	70-79%	60-69%	50-59%	0-49%

5. Experiments

5.1 Overview

Gradient boosting, a composite machine learning algorithm was the regressor adopted in the Reports Model, or the baseline, since it showed to have yielded the best results. However, in the Students Model, XGBoost was employed instead. XGBoost is a more efficient and scalable implementation of Gradient Boosting machines (Chen et al., 2015). A brief summary of the main differences between the highest performing Reports and Students Models is displayed in Table 3.

Table 3. *Comparison between the Characteristics of the Reports Model and the Students Model*

	Reports Model	Students Model
Vector Elements	Report-based	Student-based
FS_1	Teachers' comments in one lesson + understanding score + attitude score	Number of classes + understanding score quartiles + all teachers' comments
FS_2	Regular score + gender	Regular score
Text Representation	BERT	Bag of Words
Regressor	GradientBoosting	XGBoost
Total Score Estimation PTA_0	0.62	0.68

All experiments were evaluated using group 10-fold cross validation. One of the main advantages of group k-fold cross validation is that all data is used for training and testing and each instance is used once for testing. As a result, this validation method is often used especially in situations where data is limited. The average MAE, PTA_0 and PTA_1 of all ten folds were computed for all tests. As shown in Table 3, the Students Model improved the Reports Model on PTA_0 by 6%.

5.2 Results

We ran the model with all the three feature sets as described in Section 4.2. Table 4 shows the results of the Students Model. Bold values indicate the leading scores for each metric in each subject. It can be easily observed that FS3 regularly outperforms the other feature sets. In addition, from the table and as depicted in Figures 6 and 7, we can also realize that experiments using FS_1 achieved lower MAE and higher PTA_0 than FS_2 in all subjects. This strongly suggests that teachers' reports when used to predict student grades can as a matter of fact exceed the performance of a model that uses students' regular scores as the main explanatory variable.

Simultaneously, we conducted experiments for the Students Model using BERT embeddings since they proved to obtain a more accurate Reports Model. Results of the experiment using FS_3 in terms of MAE are shown in Table 5. BERT embeddings were utilized in two main ways: BERT1 and BERT2. In BERT1, the teachers' comments for each class were encoded and then summed up to be combined with the rest of the features. In a different manner, in BERT2, all comments from each class were first concatenated and then represented using BERT embeddings. However, as seen in the table, representing the teachers’ comments using bag-of-words consistently outperformed BERT embeddings in the Students Model.

Table 4. *Evaluation Metric Scores of All Subjects using the 3 Feature Sets in the Student Model. Values in Bold Indicate the Best Metric Value in its Corresponding Subject.*

	Japanese			Math			Science		
	MAE	PTA$_0$	PTA$_1$	MAE	PTA$_0$	PTA$_1$	MAE	PTA$_0$	PTA$_1$
FS$_2$	11.87	0.48	**0.45**	18.18	0.34	0.47	23.53	0.21	0.47
FS$_1$	10.62	0.54	0.42	12.42	0.47	0.46	17.32	0.34	0.48
FS$_3$	**9.42**	**0.55**	0.43	**10.32**	**0.49**	**0.48**	**12.15**	**0.47**	**0.48**
	Social Studies			English			Total		
	MAE	PTA$_0$	PTA$_1$	MAE	PTA$_0$	PTA$_1$	MAE	PTA$_0$	PTA$_1$
FS$_2$	18.62	0.31	0.48	19.28	0.29	0.49	76.67	0.36	**0.47**
FS$_1$	14.35	0.36	**0.53**	12.65	0.42	**0.5**	48.02	0.59	0.39
FS$_3$	**10.92**	**0.54**	0.42	**10.1**	**0.53**	0.42	**31.67**	**0.68**	0.32

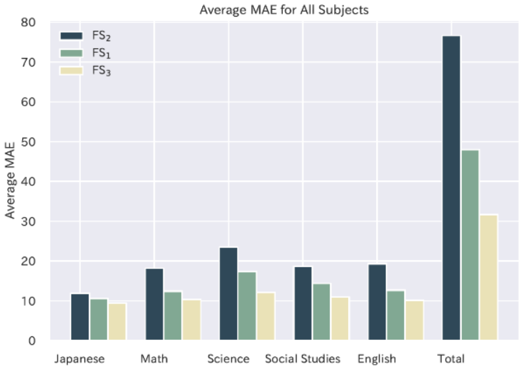


Figure 6. Average MAE for all Subjects using the 3 Feature Sets.

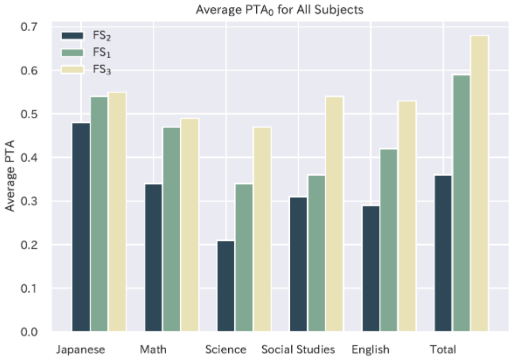


Figure 7. Average PTA$_0$ for all Subjects using the 3 Feature Sets.

Table 5. *MAE of Students Model with BERT embeddings **compared with Bag of** Words using FS$_3$*

	Japanese	Math	Science	Social Studies	English	Total
BERT1	9.52	10.93	12.75	11.13	11.07	34.30
BERT2	9.47	10.73	12.83	11.62	10.87	33.65
BoW	9.42	10.32	12.15	10.92	10.10	31.67

5.3 Compensation Experiments

Supplementary experiments were held to attempt to estimate subject scores for those students that did not attend lessons for all subjects in the cram school and thus reports were unavailable. The estimates were made based on the students’ regular scores of all the subjects. In other words, in cases where FS$_3$ was unattainable in specific subjects, the predictions were supplemented by using FS$_2$. When using this compensation for the 413 students, the total score estimation produces an MAE of 38.33 which is an acceptable value compared to the results shown in the previous section.

6. Discussion

An abridged version of the results of the baseline Reports Model is shown in Table 6 where only the BERT results are shown since they outperformed the alternative TFIDF text representation. Predicting the total score using either FS$_1$ or FS$_3$ in the Students Model yielded both lower MAE and higher PTA$_0$

and PTA$_1$ than in the Reports Model as shown in Table 6. This highlights the fact that student performance cannot be easily judged by separate reports and that an encapsulating statistical approach can accomplish a more improved performance. It is also interesting to highlight the PTA$_1$ results achieved by the Students Model where it significantly outperforms the baseline in both feature sets. Hence, it can be suggested that the Students Model is a more reliable one since it mostly either correctly identifies the letter grade or misses it by one tick. In the Students Model, PTA$_0$ + PTA$_1$ reaches an accuracy of 0.98 with FS$_1$.

On the other hand, the baseline exceeded in performance when using FS$_2$. However, this can be attributed to the fact that a different strategy or approach was taken. In the baseline, instances were report based and not student based. Thus, an effect similar to oversampling was achieved when using the second feature set.

Table 6. *Students Model Performance compared to the Baseline Model (Reports Model). Best metric values in each subject are shown in boldface, where for pta$_1$, pta$_0$+pta$_1$ is considered.*

7.

		Baseline			Students Model		
		MAE	PTA$_0$	PTA$_1$	MAE	PTA$_0$	PTA$_1$
FS$_1$	Japanese	9.47	0.27	0.23	10.62	0.54	0.42
	Math	12.36	0.45	0.07	12.42	0.47	0.46
	Science	16.66	0.40	0.11	17.32	0.34	0.48
	Social Studies	13.92	**0.55**	0.02	14.35	0.36	0.53
	English	14.51	0.52	0.02	12.65	0.42	0.5
	Total	52.02	0.49	0.07	48.02	0.59	0.39
FS$_3$	Japanese	**9.32**	0.37	0.22	9.42	**0.55**	**0.43**
	Math	**10.12**	**0.52**	0.14	10.32	0.49	**0.48**
	Science	13.31	0.43	0.18	**12.15**	**0.47**	**0.48**
	Social Studies	12.00	0.53	0.095	**10.92**	0.54	**0.42**
	English	10.99	**0.62**	0.11	**10.1**	0.53	**0.42**
	Total	33.29	0.62	0.17	**31.67**	**0.68**	**0.32**

7. Conclusion

At educational institutes where teachers observe their students closely, large amounts of unstructured data are available in the form of comments and reports. In this paper, we proposed an improved model that employs and takes advantage of these comments to produce a reliable and accurate grade prediction model. The model uses teacher observation comments taken after lessons at a cram school in Fukuoka, Japan. The comments in our previous model and the improved model are represented using different natural language processing representations from bag of words to BERT embeddings. We employed three main feature sets: teacher-related features, student-related features, and a concatenation of the two to compare and demonstrate the effect of using each feature set. Our experimental results show that using teachers' comments can not only contribute to the accuracy of existing grade prediction models based on student-related features, but they can also exceed their performance. The improved model introduced in this paper can accurately predict a students' letter grade with an accuracy of 68%, which is an increase of 6% over the previously proposed model.

However, there remains much room for improvement in our experiments. We are working on further improving the performance in hope that with such a well-defined grade prediction model, we can contribute to guiding young students by providing a more focused and personalized education to them.

Acknowledgements

This work was supported in part by e-sia Corporation and by Grant-in-Aid for Scientific Research proposal numbers (JP21H00907, JP20H01728, JP20H04300, JP19KK0257, JP18K18656). We would like to express our deepest gratitude.

References

Bretana, N. A., Robati, M., Rawat, A., Pandey, A., Khatri, S., Kaushal, K., ... & Abadia, R. Predicting Student Success for Programming Courses in a Fully Online Learning Environment. In 28th International Conference on Computers in Education, ICCE 2020 (pp. 47-56). Asia-Pacific Society for Computers in Education.

Sweeney, M., Lester, J., & Rangwala, H. (2015, October). Next-term student grade prediction. In 2015 IEEE International Conference on Big Data (Big Data) (pp. 970-975). IEEE.

Kerrey, R. (2018). Building the intentional university: Minerva and the future of higher education. MIT Press.

Han, Y., & Xu, Y. (2020). The development of student feedback literacy: the influences of teacher feedback on peer feedback. Assessment & Evaluation in Higher Education, 45(5), 680-696.

Fateen, M. & Mine, T. (2021). Predicting Student Performance Using Teacher Observation Reports. In Proceedings of The 14th International Conference on Educational Data Mining (EDM 2021) (pp. 481-486).

Lowe, R. J. (2015). Cram schools in Japan: The need for research. The Language Teacher, 39(1), 26-31.

Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. IEEE Access, 5, 8869-8879.

Rushdi, M. A., Rushdi, A. A., Dief, T. N., Halawa, A. M., Yoshida, S., & Schmehl, R. (2020). Power prediction of airborne wind energy systems using multivariate machine learning. Energies, 13(9), 2367.

Makhlouf, J., & Mine, T. (2020, November). Automatic Feedback Models to Students Freely Written Comments. In 28th International Conference on Computers in Education, ICCE 2020 (pp. 336-341). Asia-Pacific Society for Computers in Education.

Ma, H., Wang, X., Hou, J., & Lu, Y. (2017, August). Course recommendation based on semantic similarity analysis. In 2017 3rd IEEE International Conference on Control Science and Systems Engineering (ICCSSE) (pp. 638-641). IEEE.

Morsy, S., & Karypis, G. (2017, June). Cumulative knowledge-based regression models for next-term grade prediction. In Proceedings of the 2017 SIAM International Conference on Data Mining (pp. 552-560). Society for Industrial and Applied Mathematics.

Bydžovská, Hana. "A Comparative Analysis of Techniques for Predicting Student Performance." International Educational Data Mining Society (2016).

Robinson, C., Yeomans, M., Reich, J., Hulleman, C., & Gehlbach, H. (2016, April). Forecasting student achievement in MOOCs with natural language processing. In Proceedings of the sixth international conference on learning analytics & knowledge (pp. 383-387).

Luo, J., Sorour, S. E., Goda, K., & Mine, T. (2015). Predicting Student Grade Based on Free-Style Comments Using Word2Vec and ANN by Considering Prediction Results Obtained in Consecutive Lessons. International Educational Data Mining Society.

Goda, K., & Mine, T. (2011, September). Analysis of students’ learning activities through quantifying time-series comments. In International conference on knowledge-based and intelligent information and engineering systems (pp. 154-164). Springer, Berlin, Heidelberg.

Jayaraman, J. D. (2020). Predicting Student Dropout by Mining Advisor Notes. In Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020) (pp. 629-632).

Kudo, T. (2006). Mecab: Yet another part-of-speech and morphological analyzer. http://mecab. sourceforge. jp.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

cl-tohoku (2019, November 6) https://github.com/cl-tohoku/bert-japanese/tree/v1.0

Polyzou, A., & Karypis, G. (2016). Grade prediction with models specific to students and courses. International Journal of Data Science and Analytics, 2(3), 159-171.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., & Cho, H. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4).

Authoring Tool for Semi-automatic Generation of Task-Oriented Dialogue Scenarios

Emmanuel AYEDOUN[a*], Yuki HAYASHI[b] & Kazuhisa SETA[b]
[a]*Faculty of Engineering Science, Kansai University, Japan*
[b]*Graduate School of Humanities and Sustainable System Sciences, Osaka Prefecture University, Japan*
*emay@kansai-u.ac.jp

Abstract: The lack of suitable conversation opportunities is often pointed out as a major factor inhibiting second language learners' willingness to communicate in the target language. Although computer-based conversational environments have been advocated as a promising approach to mitigate this issue, high authoring costs still prevent their widespread adoption. In this paper, we present a dialogue scenario authoring system that could facilitate the rapid implementation of desirable situational dialogue scenarios, thereby lowering the dialogue scenario authoring barrier for non-programmers or even educators. To this extent, we exploit the common underlying structure of services (restaurant, hotel, travel-planning, etc.) that seem to share a certain degree of similarity at the task slevel and built a versatile dialogue scenario authoring interface that enables semi-automatic generation of services-related dialogue scenarios. Here, we describe the features of the proposed system, and present the results of a pilot evaluation study that hint on the meaningfulness of our approach towards facilitating dialogue scenarios authoring by people who do not have any previous experience designing dialogue systems components.

Keywords: Authoring tools, dialogue scenario design, adaptive language learning, conversational agents, willingness to communicate in L2

1. Introduction

In the field of second language (L2) acquisition, researchers have reported that many learners feel genuine anxiety about performing in front of others, concluding that many classrooms do not offer learners much in the way of communicative practice (Reinders, & Wattana, 2014). On the other hand, it was reported that computer-supported dialogue environments could be effective towards providing L2 learners with realistic opportunities to simulate daily conversations, whereby to alleviate emotional variables that inhibit learners' motivation towards communication in L2.

For instance, we developed an embodied conversational agent (CEWill) that provides second language learners with task-oriented spoken dialogue simulation opportunities in a restaurant context (Ayedoun, Hayashi, & Seta, 2019). Dialogue scenarios in CEWill were designed following a knowledge-based approach enabling it to achieve a deeper level of understanding and control of the conversation flow, increasing the degree of reality of interactions. However, the significant level of knowledge engineering effort and the degree of dialogue expertise that is necessary for implementing new scenarios in this system constitute an important obstacle that may limit its adoption and frequent use in real educational settings.

To alleviate such issues and promote the availability of a rich pool of realistic dialogue scenarios for second language learners, our goal is to build a dialogue scenario authoring tool that could facilitate the rapid implementation of desirable dialogue scenarios and lowering the authoring barrier for non-programmers or educators, who are not necessarily knowledge or software engineers. To this extent, we exploit the common underlying structure of services (restaurant, hotel, travel-planning, etc.) that seem to share a certain degree of similarity at the task level, and build a dialogue scenario authoring interface that enables semi-automatic generation of dialogue scenarios across various services domains.

In the present study, we present the developed dialogue scenario authoring tool, describe some of its features, and report on the meaningfulness of our approach towards reducing the authoring barrier for people who are not necessarily familiar with dialogue systems or dialogue scenario design.

2. Related Works

2.1 Authoring Tools and Learning Support

Several recent reviews have noted the effectiveness of learning support systems and particularly intelligent tutoring systems, highlighting that, well-designed systems can successfully complement or substitute other instructional models in many common academic subjects (du Boulay, 2016; Ma, Adesope, Nesbit, & Liu, 2014; Van Lehn, 2011). However, these tutoring systems remain hard to author. Hence, for the past years, extensive work has been conducted on developing authoring tools to speed up the development of learning support systems, reduce implementation workload, and lower the skill requirements. As a result, several authoring tools such as ASPIRE (Mitrovic, Martin, Suraweera, Zakharov, Milik, Holland, & Mcguigan, 2009), ASTUS (Paquette, Lebeau, Beaulieu, & Mayers, 2015), AutoTutor tools (Nye, Graesser, & Hu, 2014), SimStudent (Matsuda, Cohen, & Koedinger, 2015) have been proposed and most of them do not require advanced programming.

Authoring tools for conversation-based learning environments have focused on assisting non-technical users in the creation of pedagogical agent dialogues. AutoTutor (Graesser, Chipman, Haynes, & Olney, 2005) provides multi-agent conversational interactions to tutor students using the discourse patterns of a human tutor, and has been used across multiple domains including computer literacy and physics. To facilitate the application of AutoTutor to other domains, authoring tools have been developed to aid subject matter experts in creating dialogue-based tutors, such as the AutoTutor Script Authoring (Susarla, Adcock, Van Eck, Moreno, & Graesser, 2003) and AutoLearn (Preuss, Garc, Boullosa, 2010). Similarly, an authoring tool has been created for the Tactical Language and Culture Training System (TLCTS) that allows subject matter experts to create pedagogical dialogue for a foreign language learning training system at reduced costs (Meron, Valente, & Johnson, 2007).

However, despite the potential for increased student engagement and the reduced cost of creating lifelike virtual characters, pedagogical agents have not yet achieved widespread adoption in computer-based learning environments (Lester, Mott, Rowe, & Taylor, 2015). The available authoring tools environments, although certainly useful to implement pedagogical agents for specific domains, still seem to suffer from a lack or limited level of abstraction or versatility of their encapsulated initial domain knowledge, which limits the reusability of their key components across different domains.

2.2 Task-Oriented Dialogue Systems and Second Language Communication

The purpose and promise of computer-supported language learning technologies are to facilitate instruction that is personalized to the needs of individual learners (Kerr, 2016). Such systems have been found to be useful in engaging the learner in the educational experience (Conlan, O’Keeffe, Brady, & Wade, 2007). To sustainably enhance L2 learners’ willingness to communicate, previous research has emphasized the importance to provide learners with various realistic opportunities to simulate conversation using the target language.

Interestingly, it has been suggested that task-oriented dialogue systems, where a task should be accomplished in the target language, have a clear potential for placing the student in a realistic dialogue situation (Raux, & Eskenazi, 2004). Building on such views, in our previous works, we proposed CEWill, an embodied conversational agent that provides second language learners with opportunities to freely simulate spoken dialogue in realistic daily-life settings such as talking with a waiter in a restaurant (Ayedoun, Hayashi, and Seta, 2019). The system, which interface is shown in Figure 1, required a carefully handcrafted dialogue scenario for each situation and was equipped with a set of domain-independent conversational strategies aiming to foster the system’s ability to carry on smooth and warm interactions with learners. Results of experimental evaluations provided insights on the meaningfulness of such simulation environment, especially in countries where English learning focuses

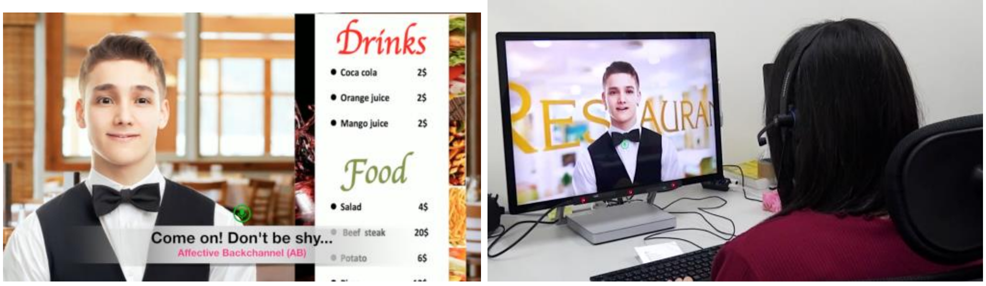


Figure 1. CEWill interface and learner interacting with the agent Peter in restaurant scenario (Ayedoun, Hayashi, and Seta, 2019).

less on the development of communicative skills and where learners have limited access to opportunities for using the target language in authentic settings.

However, to achieve a high degree of reality in interactions similar to what learners are likely to experience in daily face-to-face situations, we have also suggested that a careful design of the different dialogue scenarios would be paramount. Yet, it seems important to bear in mind that the design and implementation of a dialogue scenario are not an easy undertaking that requires a certain degree of expertise and knowledge about scenario design and dialogue systems. Hence, this important requirement could potentially constitute a barrier that may prevent the large diffusion and adoption of dialogue simulation environments such as CEWill.

3. Research Objective and Requirements

3.1 Pertinence of Issue and Research Objective

Conversational systems that provide authentic interactions simulation opportunities might be particularly beneficial for second language learners in terms of enhancing both their cognitive and emotional readiness towards communication. Yet, the widespread diffusion of such environments is tempered with the relatively important number of skills and resources required for their implementation. To address such an issue, some dialogue authoring frameworks have been developed in academia (Bohus, & Rudnicky, 2009; Lison, & Kennington, 2016). However, designing dialogue scenarios for various situations using these tools remains a challenge due to a lack or low level of reusability of their components across various domains. Murray (1999) hinted at the challenging trade-off issue related to the extent to which the difficult task of authoring learning support systems could be scaffolded: ideally, a desirable authoring tool should be both specific enough to make authoring template-based, but general enough to be attractive to many educators.

In the light of the above, the present study aims to propose a flexible authoring environment that could ease the design of dialogue scenarios over a relatively wide range of different conversation situations. If achieved, such balance between ease of authoring and flexibility of the resulting tool will allow scenario authors (i.e., educators) to have better control over the specification of dialogue behaviors, which is a crucial requirement for conversational applications, especially in the field of education. Doing so would also ultimately promote the availability of a rich pool of dialogue scenarios for L2 learners, which is desirable in terms of enhancing their motivation towards communication.

To this extent, we aim to target a suitable subset of dialogue domains that share a coherent structure at the task level and exploit such common task structure to implement an authoring tool that embeds domain-independent reusable dialogue task components. Doing so will facilitate the design process of various dialogue scenarios falling under the hood of the targeted common task. In the present work, our goal is specifically to implement a system that can serve as a proof of concept of the feasibility and meaningfulness of our approach, despite the complexity of such a challenge.

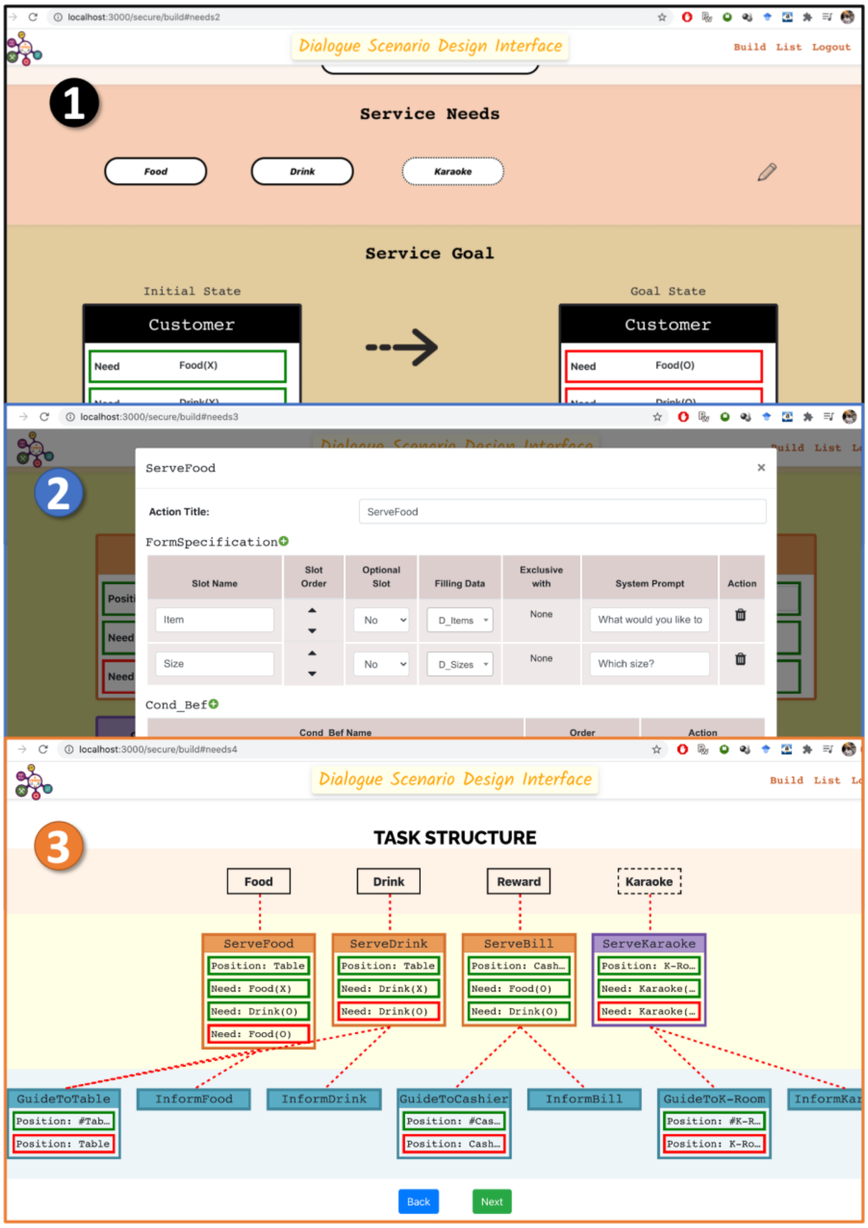
3.2 Requirements and Fulfillment

Achieving a good level of versatility while still maintaining the authoring costs at an acceptable level for content creators is a challenging task, but certainly necessary to make possible a more active and frequent use of authoring tools and learning support systems by educators who are not necessarily software engineers. Keeping this in mind and inspired by Murray (1999)'s review on authoring tools, we have identified several core requirements for the desired dialogue scenario authoring tool in the context of this study. In the following lines, we describe these prerequisites and explain how we address them in the built tool.

- **Embed a relevant level of domain-independent knowledge about task structure:** this refers to some generic knowledge about the common structure of the different dialogue domains to be targeted by the system. Such knowledge, if pre-wired and embedded in the tool, could make authoring easier and more powerful through the reuse of the same structure across various instances of dialogue domains. In such a way, dialogue scenario authors could just focus on specifying domain-specific aspects of the dialogue flow, which will significantly lessen new scenarios implementation effort and contribute to decreasing authoring time and costs. Our key idea towards covering a wide range of different dialogue domains (e.g., restaurant, hotel, transportation) is to embed a reusable domain-independent task model (services task model) in the authoring tool, as proposed in our previous work (Ayedoun, Hayashi, & Seta, 2020). Doing so could help developers with no or limited programming skills, to design and implement new service domain-related conversation scenarios in a cost-effective fashion, mostly through the specification of properties for key task components in the target domain.
- **Make possible efficient authoring flow and knowledge management:** this involves the system's ability to scaffold the dialogue scenario specification by allowing input through the use of templates, data entry forms, pop-up menus, etc. In our proposed authoring tool, whenever the range of possible input values can be limited to a finite set, scenario authors are allowed to select rather than type. Besides, we made clear separation about the different types of information that scenario authors have to deal with (i.e., actions specifications, interface parameters) and those that the system takes care of discretely behind the scene (goals and actions structure generation, dialogue flow generation, interface parameters) as we will explain in the following section. This is expected to contribute to decreasing the skill threshold for dialogue scenario modeling and allow actual educators and other people with non or less programming skills to take part in the dialogue scenario specification process.
- **Enable scenario authors to apprehend the structure of authored dialogue scenarios:** our proposed authoring tool features a user-friendly interface that allows authors to see both the static structure of the designed dialogue scenario (later shown in Figure 2(3)) and the dialogue control dynamics over possible dialogue paths (later shown in Figure 3). This enables scenario authors to interactively participate in the authoring process, examine the authored dialogue structure and make necessary refinements to achieve the desired dialogue scenario. This might especially be desirable for reducing the cognitive load associated with the design of complex dialogue scenarios.
- **Facilitate content modularity, customization, and reusability:** this refers to the authoring tool's ability to allow modular authoring of the different components needed to design a desired dialogue scenario and their storage as library structures so that they can be reused for multiple scenarios purposes or for adjusting to different learners' level. This may also facilitate the diffusion of shared dialogue scenarios design principles for studies dealing with communicative aspects of L2 acquisition. For instance, input processing, output processing, conversational strategies, etc. should be encapsulated in sub-components with well-defined interfaces that are decoupled from domain-specific dialogue flow logic. As far as this point is concerned, the dialogue scenario built with the present authoring tool is intended to be fully compatible with conversational agent systems such as CEWill so that there is no need to implement from scratch other key components dedicated to natural language processing and animation generation. Aspects related to natural language recognition and generation are thus voluntarily omitted from this paper to put more emphasis on the dialogue scenario authoring process itself.

4. Authoring Dialogue Scenarios

Figure 2 shows different windows of the built authoring tool. The system is accessible via browser and comprises the following four different windows corresponding to each stage of the authoring process: the dialogue task specification window (Figure 2(1)), the slot customization window (Figure 2(2)), the dialogue task structure visualization window (Figure 2(3)), and finally the dialogue scenario visualization window shown in Figure 3.



ServeFood

Action Title: ServeFood

FormSpecification

Slot Name	Slot Order	Optional Slot	Filling Data	Exclusive with	System Prompt	Action
Item	▲▼	No	D_Items	None	What would you like to	🗑
Size	▲▼	No	D_Sizes	None	Which size?	🗑

Cond_Bef

Cond_Bef Name	Order	Action

Figure 2. Authoring Tool Interface showing Various Windows.

In order to ease the authoring process, the system embeds a generic model of services accordingly to the conceptual framework proposed in (Ayedoun, Hayashi, & Seta, 2020), and inspired by Ferrario's prior work (Ferrario, Guarino, Janiesch, Kiemes, Oberle, & Probst, 2011). According to this framework, any service-oriented dialogue scenario can be expressed as a combination of the following three key components:

- **Core service Actions:** are those actions whose execution contributes to satisfying users' needs. In da sense, these actions characterize a service for what it is and must necessarily be exposed to the customer, e.g., for restaurant service, the action of *serving foods* or *serving drinks*;
- **Supporting Actions:** are actions necessary to the service but not explicitly mentioned as constituting the service, e.g., for restaurant service, the action *of guiding the customer to a seat* or *explaining the menu*;
- **Enhancing Actions:** are actions meant to augment the value of the service. These actions can be considered as additional services actions that are connected to but not strictly included in the main service, e.g., for restaurant service, the action of *offering karaoke*, or any other *entertainment as option*.

Along the authoring process, using the above service actions, the scenario designer specifies a hierarchically-organized service task which is similar to a type of plan structure defined in AI planning, for covering a topic. Each dialogue scenario addresses high-level goals (i.e., Needs to be satisfied) and is generated by the system following the designer's specifications as a sequence of any combination of core service actions, supporting actions, and enhancing actions. Note that the scenario designer is just required to specify execution constraints for each service action without having to care about the execution flow (i.e., how actions combine with one another), which is rather handled behind the scenes by the system. This hybrid approach is expected to consequently reduce the authoring effort by allowing scenario authors to focus solely on domain-dependent aspects of the target service domain dialogue scenario, while the authoring tool exploits the common underlying structure (i.e., service model) to manage inter-domain commonalties.

From an L2 learning support perspective, designing the dialogue scenario as a combination of these three types of service actions provides much flexibility in terms of generating dialogue content that is personalized to the needs of learners. For instance, dialogue scenarios including only core service actions (i.e., basic scenarios) may be used for beginners, while scenarios including also enhancing actions (i.e., extended scenarios) may be presented to advanced learners. That is, the dialogue system will be able to dynamically adapt the scenario content to learners' level without requiring any additional content authoring effort from scenario designers.

4.1 Dialogue Task Specification at Macro-level

The first step in the authoring process consists in specifying key components (i.e., service actions) of the targeted dialogue task structure. To begin with, the dialogue scenario author inputs customer's Need(s) to be satisfied by the target service, as shown in Figure 2(1). For example, in a restaurant scenario, *Drink* and *Food* may be set as primary Needs. According to these Need(s), the tool generates the service goal which is a representation of the initial state and goal state of the targeted service task. Based on the automatically generated service Goal, the scenario author may choose to refine both the initial and goal states by specifying some Spatio-temporal requirements or modifying the desired starting and ending state criteria for the interaction. For example, the scenario author may add a Spatio-temporal requirement (i.e., Position) for the service delivery, which can take different values at initial (e.g.: *Entrance*) and goal (e.g.: *Cashier*) states. This refinement of the service Goal enables scenario authors to clarify the big picture of the target service delivery process. At this point, the tool becomes able to generate base structures of Core and Enhancing actions to be executed to satisfy Need(s) specified in the service Goal. The suitable types of actions are automatically set according to the nature of the target Need(s). Based on the automatically generated basic structure of Core and Enhancing Actions, the scenario author may set additional Spatio-temporal requirements or constraints for each action, if necessary. Moreover, basic structures of Supporting actions are automatically generated and attached to each defined Core action and Enhancing action, as can be seen in Figure 2(3).

4.2 Dialogue Task Specification at Micro-level

Each service action may have a certain number of slots. In terms of actual dialogue flow, note that the execution of each service action can be viewed as a slot-filling driven dialogue management where dialogue slots are progressively filled through actual conversational moves between the dialogue agent and the learner. Although default slots are already predefined for convenience, the scenario author can

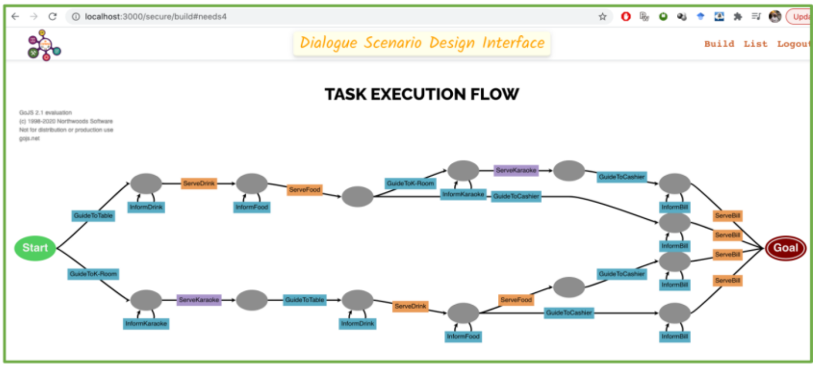


Figure 3. Authoring Tool Interface showing Automatically generated Dialogue Scenario.

still add new ones or customize exiting ones by specifying several properties that will determine the flow of slot-filling during actual dialogue. To ease the handling of this essential activity, the scenario author is prompted with a slot specification window, as illustrated in Figure 2(2).

Slots properties that are customizable include:

- **Order property:** indicates the order in which slots need to be filled to make the dialogue sound more natural. For instance, as far as the example presented in Figure 2(2) is concerned, it might be reasonable to have the slot *Item* get filled prior to the slot *Size* since different *Size* (e.g., Shot, Medium, Pitcher, Glass, Bottle) options might be available or not, depending on the type of *Item* (e.g., Beer, Wine, Tequila) that is selected.
- **Optional property:** indicates whether the slot's value is indispensable or not for the target service action execution.
- **Filling data property:** constrains the semantic type of the target slot. The configuration is conducted by selecting the appropriate type among the predefined ones. Scenario authors can still customize the existing slot types according to the restrictions of the target domain, or define new types from scratch.
- **Exclusive with property:** shows mutual exclusivity relationships between two slots. This property can be useful in cases where the filling of a given case indirectly allows the filling of another so that to avoid redundancy in slot-filling.
- **System prompt property:** shows the system prompt for triggering the learner's answer and fill the target slot. For example, when the target slot is Item under the core-service action *ServeDrink*, the system prompt might be specified as "*What would you like to drink?*" or "*Anything to drink?*".

4.3 Dialogue Task Structure Examination

The authored dialogue task structure for the target dialogue scenario is displayed as a hierarchical structure allowing the scenario author to debug and grasp the big picture of the authored domain knowledge. As illustrated in Figure 2(3), this can be seen as a static task representation of the dialogue scenario. Relations between customer Needs and service actions can be revisited by the scenario author, to ensure that intended service specification is achieved. For instance, the scenario author may decide to make appropriate revisions by adding missing constraints, or even defining new customer Needs, if necessary.

4.4 Dialogue Scenario Examination

The resulting dialogue scenario or task execution flow is automatically generated by the system, as shown in Figure 3. This can be seen as an abstract representation of the dialogue flow showing the collection of all possible task execution paths or dialogue paths with respect to constraints specified by the scenario designer through the dialogue task specification activity. The dialogue scenario is outputted by the tool in the form of a finite state machine, where each state (i.e., node) corresponds to any milestone somewhere between the initial state and the goal state; arcs (i.e., edges) connecting different states stand for service actions. Using the information displayed on this window, the scenario author can visually apprehend in which order each service action might be executed as dialogue unfold between the agent and the learner. Undesirable dialogue paths may be removed, and missing ones can be added by revisiting the specifications in the task specification windows (Figure 2(1) and 2(2)). At this point, if the author is satisfied with the generated dialogue scenario, the whole specification data can be saved for further editing or exported for integration into a conversational agent environment such as CEWill.

In the example shown in Figure 3, one can notice that the system was able to generate several dialogue scenarios for the same domain (i.e., restaurant). For instance, dialogue scenarios involving only core actions (i.e., *ServeFood*, *ServeDrink*) and corresponding supporting actions (e.g., *InformDrink*, *GuideTable*) were generated along with more extensive dialogue scenarios that also included an enhancing action (i.e., *ServeKaraoke*).

5. Pilot Study and Results

5.1 Research Questions and Experimental Settings

We conducted an experimental study to evaluate the usability of the authoring tool as well as its effectiveness in terms of reducing authoring time and preserving quality of generated dialogue scenarios. In short, the design of this preliminary experiment was guided by the necessity to investigate the following three research questions:

RQ1: Is the authoring tool accessible enough for human novices in terms of alleviating necessary knowledge engineering effort needed for dialogue scenario specification?

RQ2: Is there any noticeable difference between amount of cost needed by the system for dialogue scenarios generation compared with cost required by human novices for manual generation?

RQ3: How exhaustive are the scenarios semi-automatically generated by the system compared to those manually designed by human novices?

To answer these research questions, we designed an experimental setting to evaluate the system usability (RQ1), compare dialogue scenarios designed by human novices to those generated by the system, and see whether we could find any differences in terms of authoring time (RQ2) and quality (RQ3) of both types of dialogue scenarios. Note that the term "human novices" is used here to describe people who are not familiar with dialogue scenario design or do not have any technical experience in terms of designing dialogue systems.

The system was tested by seven undergraduate and graduate students in engineering with no previous experience authoring dialogue scenarios. We provided the participants with a short set of training materials and training tasks, which we guided them through. This required about 25 minutes. We then let them use the authoring tool to freely author the dialogue scenario they wish to implement. To ensure task homogeneity among participants, we advise them to think of dialogue in a restaurant context since it is a dialogue domain that even human novices should be reasonably familiar with.

After allowing participants to freely specify the key components of their dialogue task structure (contents of sections 4.1 to 4.3) using the tool, we asked them to stop the authoring process one step before checking the generated dialogue scenario (section 4.4). Then, at this point, to elucidate RQ2, we asked them to hand-design a dialogue scenario that would satisfy the constraints of their dialogue task structure using finite state machine symbols. We gave them as much time they needed and allowed them to check the task structure they priorly built on the interface if necessary. After they were done with this, we allow them to look at the dialogue scenario generated by the system. Then, we administrated a survey questioning their perceived difficulty of the dialogue scenario hand-design task, to elucidate RQ1, as well as their opinions on differences between their dialogue scenario and the one generated by the system, in order to investigate RQ3.

5.2 Results

RQ1: Is the authoring tool accessible enough for human novices in terms of alleviating necessary knowledge engineering effort needed for dialogue scenario specification?

In terms of the level of perceived task difficulty when designing dialogue scenarios, most participants (6 out of 7) reported that they found it difficult or very difficult to think of several dialogue paths when designing the dialogue scenario on their own (i.e., manually). On the other hand, they found that the authoring task was easier when using the authoring tool.

RQ2: Is there any noticeable difference between amount of cost needed by the system for dialogue scenarios generation compared with cost required by human novices for manual generation?

The above result was also corroborated by the amount of time spent on task. After specifying the target dialogue scenario' constraints, participants took on average about 35 minutes (M= 35.14, SD: 4.22) to come up with relatively simple dialogue scenarios. Note that since the scenario constraints were defined beforehand, the authoring tool was able to instantly (less than 1 second) generate the corresponding dialogue scenario. In other words, the system was in average, roughly more than 2100 (35 min * 60 sec) times, faster than human novices in outputting the corresponding dialogue scenario.

RQ3: How exhaustive are the scenarios semi-automatically generated by the system compared to those manually designed by human novices?

As far as differences between both dialogue scenarios are concerned, we found that valid dialogue scenarios generated by the system contained on average about 29% more edges (M=29.06, SD=8.60) than the participants' ones. This suggests that the tool was in most cases, able to generate more exhaustive dialogue scenarios than the participants. This was further corroborated by the survey results, as most participants (5 out of 7) found the dialogue scenario generated by the system completer and more exhaustive than their own. Some even mentioned that the system's dialogue scenario included some dialogue paths they have not thought of beforehand. Interestingly, two participants reported that they could not find any particular difference between theirs and the system's ones.

5.3 Discussion and Limitations

The results reported above seem to suggest that our proposed tool may substantially decrease the authoring difficulty (RQ1) involved in designing service-related dialogue scenarios. Provided that constraints on the dialogue task structure have been well-specified, even in the worst case, the system was able to generate dialogue scenarios that are at least, as exhaustive (RQ3) as those designed by human novices. In addition, the system obviously outperformed human novices in terms of time required for scenario generation (RQ2). It follows that, positive insights were obtained for each of our three research questions stated earlier in section 5.1. From these results, we hypothesise that the dialogue task specification activity which was designed for guiding the tool users step-by-step through the authoring process might have been quite beneficial towards reducing the knowledge engineering effort involved in dialogue scenario design. In addition, through the task structure and the dialogue scenario examination, experiment participants may have found it easier to keep track of their work.

These premises can be viewed as a proof of concept, suggesting that it is feasible to allow people who are not even familiar with dialogue systems, such as second language educators for example, to get involved in the design process of intelligent conversational systems. Seen in this perspective, obtained results are an important milestone towards providing second language learners computer-supported realistic opportunities to simulate various conversation situations, practice their communicative skills and reduce their apprehension towards communication in the target language.

It also seems important to emphasize that these results further deserve credit in the sense that they hint on the feasibility of building authoring frameworks that could serve as a gateway for making accessible learning support system research' outcomes to educators, so that findings and innovations from research laboratories actually reach classrooms and actual learners.

Nevertheless, although our experiment has produced some promising results, we are aware that more work is still needed to confirm the effectiveness of the proposed tool. On the one hand, the authoring tool has to be tested by actual educators to increase the quality of the user experience with its interface and validate its effectiveness. On a more conceptual level, we acknowledge that further work might be required for targeting dialogue situations which do not fall under the hood of service domains.

6. Conclusion and Future Research Directions

Authoring tools are necessary to support the more rapid delivery of computer-based learning supporting systems. However, building an authoring tool that is easy to use and not too domain specific remains extremely difficult. In this paper, we discussed on requirements that should ideally be fulfilled by a suitable authoring tool, and indicated how these prerequisites can be addressed. Then, we presented an authoring interface that enables semi-automatic generation of services related task-oriented dialogue scenarios for the field of second language learning. An experimental evaluation suggested that the proposed system could lower the dialogue scenario authoring barrier for people with no prior experience with dialogue systems, hinting on the meaningfulness of our approach.

Directions for future works include the implementation of features towards further reducing dialogue task specification workload by making available more built-in ready-to-use dialogue components. We will also devise features towards facilitating smooth and flexible integration of the dialogue scenario design module into other key modules of conversational agents such as CEWill. Finally, evaluation experiments will be conducted in real classroom settings to further understand the implications of using the authoring tool to support the tool's target (i.e., educators). More evidence will also be collected to understand how such authoring tool may impact students' learning in general and second language acquisition in particular.

References

Ayedoun, E., Hayashi, Y., & Seta, K. (2019). Adding communicative and affective strategies to an embodied conversational agent to enhance second language learners' willingness to communicate. *International Journal of Artificial Intelligence in Education, 29*(1), 29-57.

Ayedoun, E., Hayashi, Y., & Seta, K. (2020). Services Task Model Based Dialogue Scenarios Design Towards L2 WTC Support Oriented Dialogues Authoring Tool. *In International Conference on Human-Computer Interaction*, 145-163.

Bohus, D., & Rudnicky, A. I. (2009). The ravenclaw dialog management framework: Architecture and systems. *Computer Speech & Language*, *23*(3), 332-361.

Conlan, O., O'Keeffe, I., Brady, A., & Wade, V. (2007). Principles for designing activity-based personalized eLearning. In *IEEE International Conference on Advanced Learning Technologies*, 642-644.

du Boulay, B. (2016). Recent meta-reviews and meta-analyses of AIED systems. *International Journal of Artificial Intelligence in Education, 26*(1), 536-537.

Ferrario, R., Guarino, N., Janiesch, C., Kiemes, T., Oberle, D. & Probst, F. (2011). Toward an ontological foundation of services science: The general service model. In *10th International Conference on Wirtschaftsinformatik*, *2*, 675-684.

Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education, 48*(4), 612-618.

Kerr, P. (2016). Personalization of language learning through adaptive technology: Part of the *Cambridge Papers in ELT series*. Cambridge: Cambridge University Press.

Lester, L., Mott, B., Rowe, J., & Taylor, R. (2015). Design principles for pedagogical agent authoring tools. In R. Sottilare, A. Graesser, X. Hu, and K. Brawner (Eds.) *Design Recommendations for Intelligent Tutoring Systems: Volume 3 - Authoring Tools and Expert Modeling Techniques*, Orlando, FL: U.S.

Lison, P., & Kennington, C. (2016). OpenDial: A toolkit for developing spoken dialogue systems with probabilistic rules. In *Proceedings of ACL-2016 System Demonstrations*, 67-72.

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*(4), 901-918.

Matsuda, N., Cohen, W. W., & Koedinger, K. R. (2015). Teaching the teacher: Tutoring SimStudent leads to more effective Cognitive Tutor authoring. *International Journal of Artificial Intelligence in Education, 25*(1), 1-34.

Meron, J., Valente, A., & Johnson, W. L. (2007). Improving the authoring of foreign language interactive lessons in the Tactical Language Training System. In *Workshop on Speech and Language Technology in Education,* 33-36.

Mitrovic, A., Martin, B., Suraweera, P., Zakharov, K., Milik, N., Holland, J., & Mcguigan, N. (2009). ASPIRE: An authoring system and deployment environment for constraint-based tutors. *International Journal of Artificial Intelligence in Education, 19(2)*, 155-188.

Murray, T. (1999): Authoring intelligent tutoring systems: an analysis of the state of the art. *International Journal of Artificial Intelligence in Education, 10*, 98-129.

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education, 24*(4), 427-469.

Paquette, L., Lebeau, J. F., Beaulieu, G., & Mayers, A. (2015). Designing a knowledge representation approach for the generation of pedagogical interventions by MTTs. *International Journal of Artificial Intelligence in Education, 25*(1), 118-156.

Preuss, S., Garc, D. & Boullosa, J. (2010). AutoLearn's authoring tool: A piece of cake for teachers. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, 19-27.

Raux, A., & Eskenazi, M. (2004). Using task-oriented spoken dialogue systems for language learning: potential, practical applications and challenges. In *InSTIL/ICALL Symposium*.

Reinders, H., & Wattana, S. (2014). Can I say something? The effects of digital game play on willingness to communicate. *Language Learning & Technology, 18*(2), 101-123.

Susarla, S., Adcock, A., Van Eck, R., Moreno, K., Graesser, A. C., & The Tutoring Research Group (2003). Development and evaluation of a lesson authoring tool for AutoTutor. In *AIED2003 Supplemental Proceedings*, 378-387.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197-221.

Diverse Linguistic Features for Assessing Reading Difficulty of Educational Filipino Texts

Joseph Marvin IMPERIAL[ab*] **& Ethel ONG**[b]
[a]*National University, Manila, Philippines*
[b]*De La Salle University, Manila, Philippines*
**jrimperial@national-u.edu.ph*

Abstract: To ensure quality and effective learning, fluency, and comprehension, the proper identification of the difficulty levels of reading materials should be observed. In this paper, we describe the development of automatic machine learning-based readability assessment models for educational Filipino texts using the most diverse set of linguistic features for the language. Results show that using a Random Forest model obtained a high performance of 62.7% in terms of accuracy, and 66.1% when using the optimal combination of feature sets consisting of traditional and syllable pattern-based predictors.

Keywords: Readability assessment, Filipino, linguistic features, natural language processing

1. Introduction

Reading is a vital life skill that children learn early during their primary years. It is key towards gaining good academic performance, as the ability to read not only affects language and literacy development, but also the ability to understand and solve math problems and comprehend science facts. Learning to read entails multiple processes that include word recognition (Imperial & Ong, 2021), comprehension, and fluency in order to make meaning from the printed text (Klauda & Guthrie, 2008). Thus, the proper identification of readability levels of children's literature is very important. Previous studies have reported that children who read educational materials such as story and picture books often experience frustration and boredom when the reading level is not within their range of ability (Gutierrez, 2015; Guevarra, 2011). However, readability assessment comes with a challenge—manual extraction of linguistic variables such as determining occurrences of polysyllabic words, calculating average character length, tagging grammatical labels of words and plugging them to handcrafted formulas can be heavily time-consuming and tedious. In addition, linguistic variables that can potentially affect the difficulty levels of texts vary from language to language. In this study, we investigate the performance of other possible factors affecting readability from a language's characteristics such as lexical, language model, syllabication rules, and morphology for readability assessment of educational text in Filipino.

2. Related Work

While high-resource languages such as English (Vajjala & Lučić, 2018; Collins-Thompson & Callan, 2014; Feng et al., 2010; Collins-Thompson & Callan, 2005) and German (Hancke & Meurers, 2013; Hancke et al., 2012) boast a rich history of research efforts in readability assessment, the same cannot be said for low-resource languages where machine-readable data is limited and NLP-based tools for extracting complex linguistic predictors are yet to be developed. To compare, research on high-resource languages have explored more potential linguistic predictors, even extending to psycholinguistics (Vajjala & Meurers, 2014), cognitive-based features (Crossley et al., 2008) and gaze movement (Gonzalez-Garduno & Søgaard., 2017), than low-resource languages such as Filipino, Bangla, Telugu, and other South Asian languages.

In Filipino, experts conducting research in the field such as Macahilig (2015) and Gutierrez (2015) have outlined possible linguistic factors that should be considered specifically for the language. These factors include the lexical categories of words, syllable patterns, syntactic features, and text structure. To date, few research works have explored using these prescribed features. Guevarra (2011) initiated the use of a machine learning model via logistic regression with seven traditional features such as number of unique words, average number of syllables, and total number of sentences. Later works by Imperial et al. (2019), and Imperial and Ong (2020, 2021) explored the inclusion of lexical features such as lexical densities, foreign and compound word densities, as well as language model features such as word and character trigrams and noted significant improvement to performance when using the latter. We further extend these works by supplying linguistic predictors that have not been explored yet such as features on syllable patterns based on the orthography of the Filipino language and morphological features which have never been used before in Filipino text readability research, thus, completing the recipe for text-based features for readability assessment of Filipino texts.

3. Educational Filipino Text Corpus

The Filipino text corpus used for training and validating our ML models is derived from two sources: *Adarna House Corpus* and *DepEd Commons*. Our study focuses only on resources for early grade learners as this is the group where reading materials are abundant online.

Adarna House Corpus. We obtained permission from Adarna House Inc., the largest children's literature publisher in the Philippines accredited by the Department of Education, to use their machine-readable copy of leveled reading materials. A total of 174 reading materials were obtained, spanning from grades 1 to 3 in the form of fictional story books written in Filipino. This resource has been pre-annotated by Adarna's in-house language experts and writer-researchers in terms of readability level.

DepEd Commons. DepEd Commons is an online platform launched by the Department of Education (DepEd) which contains open-source books and reading materials for all basic education subjects in various grade levels such as *English, Filipino, Science, Araling Panlipunan* (Social Studies), *Arts*, and *Physical Education*. These are available for free download by public and private school teachers and learners. A total of 91 reading passages in the form of short stories also spanning from grades 1 to 3 were extracted from Filipino activity books obtained from the repository and added to the current data count with the Adarna House.

4. Diverse Linguistic Features

The concept of automating the readability assessment task draws from the notion of extracting a wide range of linguistic features often recommended by experts that can potentially affect the readability of texts (Macahilig, 2015; Gutierrez, 2015). As such, we performed extraction of linguistic feature sets from the corpus as numerical vector representations required for model training. A total of **54 linguistic predictors** from 5 different feature sets covering various facets of texts such as **surface-based or traditional (TRAD), lexical (LEX), language structure (LM), syllable pattern (SYLL)**, and **morphological predictors (MORPH)** were used for this study. To note, this is the most extensive number of linguistic factors considered for Filipino readability assessment to date, much more than features extracted from the works of Macahalig (2015), Imperial et al. (2019), and Imperial and Ong (2020) with 15, 15, and 25 predictors respectively. Table 1 describes each predictor from the extracted feature sets.

Table 1. *Breakdown of Various Linguistic Predictors per Feature Set.*

Feature Set	Count	Predictors
TRAD - Traditional or surface-	7	Total word, sentence, phrase, polysyllabic word counts.

based features based on counts and frequencies.		Average word length, sentence length, syllable per word.
LEX - Lexical or context carrying features via part-of-speech categories.	9	Type-token variations (regular, logarithmic, corrected, root). Noun and verb token ratio. Lexical density. Foreign word and compound word density.
LM - Language model features based on perplexity.	9	Language models trained on three levels (L1, L2, and L3) of the external DepEd Commons corpus using n-gram values of {1, 2, 3}.
SYLL - Syllable pattern densities based on the prescribed national orthography.	10	Consonant cluster density. Densities of the prescribed Philippine orthography on syllable patterns: {*v*, *cv*, *vc*, *cvc*, *vcc*, *cvcc*, *ccvc*, *ccvcc*, *ccvccc*} where *c* and *v* are consonant and vowel notations.
MORPH - Morphological features based on verb inflection.	19	Densities of various foci of verbs based on tense: {actor, object, benefactive, locative, instrumental, referential}. Densities of various foci of verbs based on aspect: {infinitive, perfective, imperfective, contemplative, participle, recent-past, auxiliary}.

5. Experiment Setup

To develop models for automatic readability assessment, we select commonly-used machine learning algorithms for document classification, namely Random Forest (RF), and Support Vector Machines (SVM), and use the extracted feature sets from the reading materials and their corresponding labels. During actual training, *k*-fold cross validation was performed with *k* = 5 for each algorithm. The performance of the models were evaluated using accuracy, precision, recall, and F1 score.

6. Performance Results of Machine Learning Models

We conducted a series of model training using the extracted linguistic feature sets to empirically understand and analyze each selected machine learning algorithm's performance on the automatic readability assessment task. Each feature set is selected and modeled singularly followed by combining feature sets together incrementally until all combinations are used. All models from these experiments have their hyperparameters optimized through exhaustive grid search.

Table 2. *Ablation Experiments of Model Training using the Extracted Linguistic Feature Sets for* ***Support Vector Machines*** *with Hyperparameters Optimized. The top performing combination of features is highlighted in* ***boldface****.*

Feature Set	Accuracy	Precision	Recall	F1 Score
TRAD	0.475	0.460	0.475	0.465
LEX	0.424	0.421	0.424	0.415
SYLL	0.458	0.447	0.458	0.449
LM	0.390	0.369	0.390	0.373
MORPH	0.322	0.366	0.322	0.308
TRAD + LEX	0.475	0.449	0.475	0.406
TRAD + LM	**0.542**	**0.532**	**0.542**	**0.536**
TRAD + SYLL	0.508	0.484	0.508	0.490
TRAD + MORPH	0.492	0.476	0.492	0.465
TRAD + LEX + SYLL	0.475	0.452	0.475	0.460
TRAD + LEX + LM	0.475	0.475	0.475	0.475
TRAD + LEX + MORPH	0.475	0.456	0.475	0.456
TRAD + LEX + SYLL + LM	0.492	0.494	0.492	0.492
TRAD + LEX + SYLL + MORPH	0.492	0.499	0.492	0.495

ALL	0.492	0.481	0.492	0.485

From the training results of SVM using the singular feature sets in Table 2, TRAD emerged as the top single-feature predictor amongst all other feature sets. In addition, using MORPH feature sets obtained the lowest from the five. This result further strengthens the importance of TRAD features in readability assessment as done previously. The best performing model from the Support Vector Machine experiments uses a combination of TRAD + LM feature sets with an accuracy of 0.542, precision of 0.532, recall of 0.542, and F1 score of 0.536. With LM features present in all top-performing models, one inference that can be made from this is that the language model features are often the most-used support vectors for discriminating readability levels between classes with respect to how Support Vector Machine works.

Table 3. *Ablation Experiments of Model Training using the Extracted Linguistic Feature Sets for* ***Random Forest*** *with Hyperparameters Optimized. The top performing combination of features is highlighted in* ***boldface****.*

Feature Set	Accuracy	Precision	Recall	F1 Score
TRAD	0.525	0.528	0.525	0.526
LEX	0.508	0.487	0.508	0.492
SYLL	0.525	0.530	0.525	0.515
LM	0.458	0.448	0.458	0.452
MORPH	0.475	0.494	0.475	0.475
TRAD + LEX	0.593	0.588	0.593	0.588
TRAD + LM	0.610	0.594	0.610	0.597
TRAD + SYLL	**0.661**	**0.651**	**0.661**	**0.640**
TRAD + MORPH	0.576	0.564	0.576	0.563
TRAD + LEX + SYLL	0.644	0.634	0.644	0.634
TRAD + LEX + LM	0.610	0.573	0.610	0.583
TRAD + LEX + MORPH	0.525	0.537	0.525	0.529
TRAD + LEX + SYLL + LM	0.627	0.602	0.627	0.605
TRAD + LEX + SYLL + MORPH	0.593	0.573	0.593	0.571
ALL	0.627	0.623	0.627	0.624

Using the singular feature sets for Random Forest, the TRAD emerged as one of the feature sets used by the top-performing models similar to Support Vector Machines. Thus, the efficacy of TRAD features is substantial for readability assessment of Filipino texts regardless of what machine learning algorithm is used. This result can also serve as a further basis that the use of traditional features from previous works (Villamin & de Guzman, 1979; Gutierrez, 2015; Macahilig, 2015) is practical. The other feature set that contributed towards obtaining the best performance is the SYLL feature set or predictors using syllable patterns (**TRAD + SYLL**). We infer that the syllable patterns can also substantially influence word difficulty in Filipino whereas more common patterns such as *cvc*, *vc*, and *cv* appear on highly readable texts and uncommon patterns such as *ccvcc* and *ccvccc* are more common on advanced texts. It is also of no surprise that the combination of these two feature sets would result in the top-performing model with an accuracy of 0.661, precision of 0.651, recall of 0.661, and F1 score of 0.640 among all models. For scale, the 0.661 accuracy is the current highest score for a readability assessment model in Filipino up to date from this work on studying the most number of linguistic features.

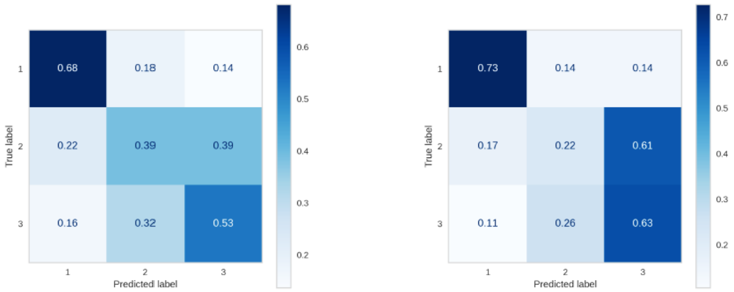


Figure 1. Confusion Matrices of Top Models using *All* 54 Extracted Features (Left) and *Optimal Combination* of Features (TRAD + LM) for Support Vector Machines (Right).

We also provide an analysis of correctly classified instances and fail cases via confusion matrices for the trained readability models. Figure 1 depicts the two confusion matrices for the model using all the feature sets (left) and the top-performing model using TRAD + LM (right) for comparison for Support Vector Machines. From the figures, both models often confuse the readability levels between Grade 2 and Grade 3. This observation is more prominent in the case of the TRAD + LM model where the shade is darker. Though it obtained the highest scores for the evaluation metrics, it does not perform well when classifying the Grades 2 and 3 text compared to the model using all feature sets. One inference from this is the decrease in the number of features (38 features not used) in TRAD + LM caused some confusion when differentiating Grade 2 from Grade 3. We note that there is a slight *trade-off* between the use of all features versus the optimal combination on the performance of models.

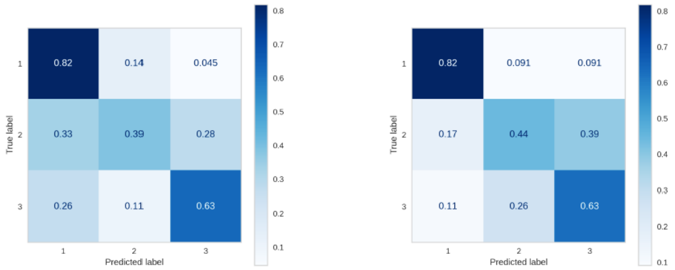


Figure 2. Confusion Matrices of Top Models using *All* 54 Extracted Features (Left) and *Optimal Combination* of Features (TRAD + SYLL) for Random Forest (Right).

Figure 2 describes the two confusion matrices for the model using all the feature sets (left) and the top-performing model using TRAD + SYLL (right) for Random Forest. From the figures, the majority of the predicted labels were matched to their corresponding true labels and both also performed better when classifying Grades 1 and Grades 3 from other grade levels. The confusion of discriminating between Grades 2 and 3 from the models developed using Support Vector Machines is not exhibited using Random Forest. This can be traced to the nature of training Random Forest using an ensemble of decision trees using random feature subsets by a number of tree estimators to avoid bias and reduce overfitting (Ho, 1995).

7. Conclusion

In order to ensure quality and effective learning and comprehension, the proper identification of the difficulty levels of reading materials should be observed. In this study, an automatic readability assessment model was trained from children's education materials using deep, hybrid linguistic feature sets spanning traditional, lexical, language model, syllable pattern, and morphological predictors. Results showed that the Random Forest algorithm outperformed Support Vector Machines in identifying readability levels of texts using a hybrid combination of traditional features (TRAD) and syllable pattern features (SYLL) with an accuracy of 0.661. Future directions of the study can focus on inclusion of non text-based features such as audio from recorded readings of texts and psycholinguistic features such as age-of-acquisition of words can be explored to achieve a multi-modal approach in building the readability assessment model.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable feedback as well as Adarna House Inc. for allowing us to use their dataset for this study.

References

Collins-Thompson, K., & Callan, J. P. (2004). A language modeling approach to predicting reading difficulty. *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*: HLT-NAACL 2004 (pp. 193-200).

Collins‐ Thompson, K., & Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13), 1448-1462.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *Tesol Quarterly*, 42(3), 475-493.

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A comparison of features for automatic readability assessment.

Gonzalez-Garduno, A. V., & Søgaard, A. (2017). Using gaze to predict text readability. *In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 438-443).

Guevarra, R. C. (2011). Development of a lipino text readability index. UP College of Engineering Professorial Chair Colloquium.

Gutierrez, M. R. M. (2015). The suitability of the fry and smog readability formulae in determining the readability of lipino texts. *The Normal Lights*, 8 (1).

Hancke, J., Vajjala, S., & Meurers, D. (2012). Readability classification for German using lexical, syntactic, and morphological features. *In Proceedings of COLING 2012* (pp. 1063-1080).

Hancke, J., & Meurers, D. (2013). Exploring CEFR classification for German based on rich linguistic modeling. *Learner Corpus Research*, 54-56.

Ho, T. K. (1995). Random decision forests. *In Proceedings of 3rd International Conference on Document Analysis and Recognition* (Vol. 1, pp. 278-282). IEEE.

Imperial, J. M., Roxas, R. E., Campos, E. M., Oandasan, J., Caraballo, R., Sabdani, F. W., & Almaroi, A. R. (2019). Developing a machine learning-based grade level classifier for Filipino children's literature. *In 2019 International Conference on Asian Language Processing (IALP)* (pp. 413-418). IEEE.

Imperial, J. M., & Ong, E. (2021). Application of Lexical Features Towards Improvement of Filipino Readability Identification of Children's Literature. *arXiv preprint* arXiv:2101.10537.

Imperial, J. M., & Ong, E. (2020). Exploring Hybrid Linguistic Feature Sets to Measure Filipino Text Readability. *In 2020 International Conference on Asian Language Processing (IALP)* (pp. 175-180). IEEE.

Imperial, J. M., & Ong, E. (2021). Semi-automatic Construction of Sight Words Dictionary for Filipino Text Readability. In Knowledge Management and Acquisition for Intelligent Systems: *17th Pacific Rim Knowledge Acquisition Workshop*, PKAW 2020, Yokohama, Japan, January 7–8, 2021, Proceedings 17 (pp. 168-177). Springer International Publishing.

Klauda, S. L., & Guthrie, J. T. (2008). Relationships of three components of reading fluency to reading comprehension. *Journal of Educational Psychology*, 100(2), 310.

Macahilig, H. B. (2015). A content-based readability formula for filipino texts. *The Normal Lights*, 8 (1).

Vajjala, S., & Meurers, D. (2014). Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *ITL-International Journal of Applied Linguistics*, 165(2), 194-222.

Vajjala, S., & Lučić, I. (2018). OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 297-304).

Villamin, A. M., & de Guzman, E. (1979). Pilipino readability formula: The derivation of a readability formula and a Pilipino word list. Language Study Center: Philippine Normal University.

Challenges to Applying Performance Factor Analysis to Existing Learning Systems

Cristina MAIER[a*], Ryan S. BAKER[b] & Steve STALZER[c]
[a]*McGraw Hill Education, USA*
[b]*University of Pennsylvania, USA*
[c]*McGraw Hill Education, USA*
*cristina.maier@mheducation.com

Abstract: The last decade has seen a wide variety of new algorithms proposed for knowledge tracing in adaptive learning. However, with the exception of Bayesian Knowledge Tracing (BKT), most of these algorithms' properties for real-world usage have not been thoroughly studied. In this paper, we consider real-world practical concerns around the scaled use of Performance Factors Analysis (PFA), another widely researched algorithm: developing models that work for skills that were rare or unavailable in initial data sets, skills encountered by many students but only in one or two items, content tagged with both common and rare skills, and whether skills are compensatory or conjunctive. We map these limitations to the problem of model degeneracy, not yet explored in detail for PFA. We discuss the scope and properties of each challenge, and then discuss potential solutions.

Keywords: Knowledge tracing, performance factor analysis, adaptive learning

1. Introduction

In recent years, a proliferation of ways to model student knowledge have emerged for adaptive learning environments. Early research focused on variants of a single algorithm, Bayesian Knowledge Tracing (BKT; Corbett & Anderson, 1995). More recently, dozens of algorithms – many of them variants on logistic regression or neural networks – have become popular. However, most commercial adaptive learning systems still continue to either use BKT (Sales & Pane, 2019) or simpler heuristics such as three-in-a-row correct. Part of the reason for this choice is the speed and ease of implementing BKT – the code to incorporate BKT into a system can fit into a single page, and there are several public packages for fitting BKT, which can run through large data sets in under a day. However, another reason is the relatively deep understanding of BKT's properties in real-world usage (e.g. Baker et al., 2008; van de Sande, 2013; Pelánek et al., 2016; Sales & Pane, 2019). For example, researchers have studied the phenomenon of degenerate parameters in BKT -- parameter values where the algorithm's behavior does not match the intuitive interpretation of what the parameters should mean (Baker et al., 2008; van de Sande, 2013). The ELO algorithm, used in many adaptive learning systems, has also been formally studied (e.g. Pelánek et al., 2016; Yudelson, 2019).

The real-world properties of neural network variants is also being studied. However, despite excellent predictive performance (e.g. Piech et al., 2015; Yeung, 2018), researchers found that the first variant (Deep Knowledge Tracing, DKT; Piech et al., 2015), had limitations for real-world usage, such as predicted performance going down after correct answers and large fluctuations in prediction from action to action (Yeung, 2018). Concerns have also been raised about the interpretability and usefulness of this model's predictions for teachers and other end users (Zhang et al., 2017). Research continues into the properties of these types of algorithms (Yeung, 2018; Lee et al., 2021).

In this paper, we consider another algorithm, Performance Factors Analysis (PFA; Pavlik et al., 2009). PFA performs competitively in predicting performance (Gong et al., 2010; Scruggs et al., 2020), though more poorly than DKT variants for large data sets (Gervet et al., 2020). However, its properties for real-world use have been less thoroughly studied.

2. Knowledge Tracing in Adaptive Learning

Knowledge Tracing is commonly posed as the problem of trying to use student performance on a set of items to predict performance on future items (e.g. Piech et al., 2015). However, some have argued instead that it should be defined as the problem of trying to use student performance on a set of items to predict latent knowledge that leads to better performance outside the learning system (Corbett & Anderson, 1995; Scruggs et al., 2020). In this section, we will define PFA and mention a few other algorithms that will be discussed within the paper.

PFA, in its typical formulation, predicts performance on a given item, at a given time, using a student's past number of successes, multiplied by a weight γ fit to each of the item's skills, a student's past number of failures, multiplied by a weight ρ fit to each of the item's skills, a weight β which, depending on the variant of PFA, is applied across all contexts, across all items linked to the current skill (the most common approach and what we will use here), across all items of the same "item-type", or for individual items**.** These features are inputted into a logistic function to obtain a prediction, p(m), between 0 and 1 for success for a given student on a given (future) item.

$$\mathbf{m(i, j \in KCs, s, f)} = \sum_{\mathbf{j \in KCs}} (\boldsymbol{\beta}_\mathbf{j} + \boldsymbol{\gamma}_\mathbf{j}\mathbf{s}_\mathbf{i,j} + \boldsymbol{\rho}_\mathbf{j}\mathbf{f}_\mathbf{i,j}); \qquad \mathbf{p(m)} = \frac{\mathbf{1}}{\mathbf{1 + e^{-m}}}$$

Within this equation, KCs are the Knowledge Components (i.e. skills, benchmarks) linked to the item. Parameter $\boldsymbol{i}$ represents the current learner. Parameters $\boldsymbol{\beta_j}, \boldsymbol{\gamma_j}$ and $\boldsymbol{\rho_j}$ represent the learned parameters for skill $\boldsymbol{j}$, $\boldsymbol{s_{i,j}}$ represents the number of successful practices made by learner $\boldsymbol{i}$ on skill $\boldsymbol{j}$ thus far, and $\boldsymbol{f_{i,j}}$ is the number of failed practices made by learner $\boldsymbol{i}$ on skill $\boldsymbol{j}$ thus far. PFA is therefore able to model learning in common real-world situations where multiple skills are associated with an item.

Despite PFA's widespread use in research and its competitive success on predicting future student performance in some studies (e.g. Gong & Beck, 2011; Gong, Beck, & Heffernan, 2010; Scruggs et al., 2020), there has been relatively little study of what factors impact PFA's behavior in real-world learning settings. This paper's goal is to study the factors that have emerged for other algorithms, to better understand the use of PFA in real-world learning. In section 3, we discuss our data set and PFA's baseline performance. In section 4, we focus on four challenges: insufficient data for a student and skill, degenerate parameters, rare benchmarks, and compensatory vs. conjunctive skill relationships. We conclude the paper with a brief discussion of other potential challenges for PFA.

3. Dataset and the Baseline PFA Results

The experiments provided within this paper use data from Reveal Math Course 1, a digital core math product for 6th grade. The platform provides instructional material, such as lessons and units, and assessments with question items. The items in this courseware are tagged with metadata that represent a set of skills. For these experiments we used a type of skill called a benchmark that corresponds to mathematics standards. Items can be tagged with multiple benchmarks. Throughout the paper the terms skill and benchmark will be used interchangeably. The data we used for the experiments come from three U.S. school districts that use NGA Center/CCSSO Common Core Standards.

We obtained data from 3073 students in 46 schools, who used the system between August 2019 to January 2021. We extracted information about scored items, and normalized the scores to values between 0 and 1. Though the data had partial credit, only 1.35% of responses received a partial score. Therefore, we only used binary scores, assigning any normalized partial score < 1 a final score of 0.

For analysis of model quality, we split our dataset into training and testing sets. We randomly selected 20% of the students (614 students; 52,516 data points; 64 benchmarks) to the testing dataset, leaving the remaining 80% of the students (2,459 students; 208,553 data points; 62 benchmarks) in the training dataset. The two benchmarks missing from the testing dataset were very rare; each had only one data point. The dataset contains a mix of multi-skill and single-skill items. 52.77% of the training set involved multi-skill items, and 53.81% of the testing set involved multi-skill items.

We analyzed our models' performance across the testing dataset. For some experiments, we analyzed the results for four subcategories of datapoints linked to common and/or rare benchmarks. We consider a benchmark common if the training dataset has at least 200 students that have at least 3 data points (practices, items) linked to that benchmark (cf. Slater & Baker, 2018), and rare if it does not satisfy this condition. 27 benchmarks were common and 37 were rare. The four testing data subcategories are: datapoints linked to at least one common benchmark (50,251, out of which 56.03% are multi-skill), datapoints linked to only common benchmarks (49,371, out of which 52.25% are multi-skill), datapoints linked to at least one rare benchmark (3,145, out of which 31.22% are multi-skill), and datapoints linked only to rare benchmarks (2,265, out of which 4.5% are multi-skill).

3.1 Baseline PFA Results

We trained a "baseline" model using the original PFA formula on all 64 benchmarks. In this model, β parameters were bounded from –3 to 3, and γ and ρ parameters were bounded from –1 to 1. Bounds were used to keep parameters within a reasonable range and speed the training process (training with no bounds had minimal impact on AUC and RMSE). Overall, AUC is in the 0.78-0.80 range (Table 1), except (unexpectedly) for rare benchmarks, where AUC reaches 0.83. These values are somewhat higher than typically seen for PFA in other papers (Gervet et al., 2020; Gong et al., 2010; Scruggs et al., 2020). RMSE values, in the 0.42-0.44 range, are more in line with values seen in past papers.

Table 1. *Baseline unmodified PFA Model - Validation Results (Testing Dataset)*

Category Data Points	AUC (bounds)	AUC (no bounds)	RMSE (bounds)	RMSE (no bounds)
All	0.7818	0.7810	0.4245	0.4207
At least one common benchmark	0.7809	0.7801	0.4244	0.4203
Only common benchmarks	0.7797	0.7787	0.4235	0.4193
At least one rare benchmark	0.7954	0.7930	0.4408	0.4420
Only rare benchmarks	0.8320	0.8302	0.4268	0.4297

4. Challenges, Potential Solutions and Experimental Results

4.1 Insufficient Number of Practices

One challenge with using PFA appears when students in the data set have insufficient practice with a skill. Effective estimation for PFA, like all algorithms, depends on having sufficient data, and sufficient data set size for knowledge tracing depends on the number of data points per student and skill as well as the overall data set size, and insufficient data set size can lead both to poorer prediction and extreme parameter values (Slater & Baker, 2018). Training a PFA model with data that does not contain enough students with a sufficient number of practices for a skill might yield unsatisfactory performance, and create degenerate or extreme learned parameters. The question of how much practice is needed to get reliable parameters has been studied for BKT (Slater & Baker, 2018), but not yet for PFA, although one comparison across data sets suggests that PFA may need less data than DKT (Gervet et al., 2020).

Using the baseline PFA approach, we ran several experiments with different thresholds for the maximum number of practices, to simulate the impact of having less data per student. Specifically, we filtered the training data points to retain only the data points that are linked to at least one skill for which the student had at most 'threshold' practices. We trained multiple models using the filtered training data points for different thresholds (2 to 7), and then we validated each model against the entire testing dataset. We found major improvement from 2 to 3 practices (AUC increased from ~0.753 to ~0.772), with continued improvement up to practice 6 (AUC ~0.78). For the trained models with fewer than 5 practices (i.e. threshold < 5), we observed more degenerate learned parameters.

We also investigated how insufficient number of practices with a skill impacted prediction, by training a PFA model using the entire training dataset, and validating across different subsets of the testing dataset. We filtered the testing dataset such that we retained only the datapoints which were linked to at least one skill for which this datapoint represented for the student a practice number less or equal to a given threshold (2-20). The higher the threshold, the more practices allowed. With more practices, the validation results improved – most substantially from the second to fifth practices (AUC increased from ~0.725 to ~0.755), followed by a slower increase up to around practice 12, after which the performance flattens out (AUC ~0.78).

4.2 Degenerate Parameters

Another challenge is degenerate learned parameters. Model degeneracy has been discussed relatively thoroughly for BKT (Baker, Corbett, & Aleven, 2008; Pardos & Heffernan, 2010; van de Sande, 2013). A model is degenerate "where parameter values violate the model's conceptual meaning (such as a student being more likely to get a correct answer if he/she does not know a skill than if he/she does)." (Baker, Corbett, & Aleven, 2008, p. 406). There have been reports of degenerate behavior for DKT as well (Yeung, 2018). We believe there are three cases where a PFA model is degenerate. First, when $\gamma < 0$ -- this indicates that if a student obtains a correct answer, they are likely to do more poorly in the future. Second, when $\gamma < \rho$ -- this indicates that a student's future performance will be better if they get the item wrong now than if they get the item right now. Third, when γ and ρ are both zero, no matter what the student does, the predictions will not change. It is worth noting that a fourth case when $\rho > 0$ -- is not degenerate, due to the multiple functions the parameters perform in PFA. In this case, the rate of learning the skill may outweigh the evidence of lack of student knowledge that an incorrect answer provides. So long as $\gamma > \rho$, a positive ρ is conceptually acceptable.

Several causes could produce degenerate PFA parameters. For example, if β is used at the skill or item type level (as in Pavlik et al., 2009), then a learning system that moves students from easier items to harder items, within the same nominal skill, will produce $\gamma < 0$. In cases where items are tagged with multiple skills, collinearity between skills could produce degenerate parameters.

In practice, within our full test data set, a baseline PFA model resulted in 6 skills (9% of skills) that had type 1 degeneracy, and 7 skills (11% of all skills) had type 3 degeneracy. No skill showed type 2 degeneracy. All degenerate learned parameters were linked to rare benchmarks**.** However, type 2 degeneracy was observed when training models that only used data from 4 or fewer practices.

This problem can be addressed as in (Corbett & Anderson, 1995), by constraining parameter values to be non-degenerate. When we do this with the Reveal Math Course 1 data and a baseline PFA model, constraining γ to be within [0,1] and ρ within [-1,0], we get slightly better test set performance than for the baseline PFA with looser bounds, with an AUC ROC of 0.7834 and an RMSE of 0.4219. However, even with these constraints, just as many (7) skills still had type 3 degeneracy.

It is noteworthy that baseline PFA had excellent AUC ROC on rare skills despite having degenerate parameters for some of those skills. This may be due to the fact that these skills generally had at most 1 or 2 practices for a student, which means that γ and ρ values either did not count at all or played a minimal role. Overall, then, it seems like degeneracy is a challenge for PFA, but primarily so when data is limited. The *merged-rare model* proposed next offers a potential solution to this problem.

4.3 Rare vs Common Skills

Another challenge is the handling of rare skills, which may occur in several situations, including when instructors frequently choose not to assign some concepts, or some items are tagged with skills that are not directly taught within the courseware, such as prerequisite skills. It is also possible that authors may under/over-tag items. Thus, there may be many cases where tags differ in granularity and frequency.

A learning system needs to be able to handle both common and rare skills. Depending on how rare a skill is, there might not be enough data points for precise parameter estimation when training for those skills, or there might not be enough data points per student. Simulation-based research has been conducted to distill guidelines for minimum data set size for BKT (Slater & Baker, 2018), but not yet for PFA, which would be more complicated due to multi-skill items. There is also the case when we have to deal with an item tagged with a skill with no past data. For both these cases, we need to decide which skills to train our model on, and then need a way to handle performance prediction for items tagged with skills the model was not trained for. Hence, we propose to adjust the original PFA formula to differentiate between rare and common skills. While different β, γ and ρ parameters are fit for common skills, a common set of parameters ($\beta_d, \gamma_d, \rho_d$) are used for rare skills:

$$\mathbf{m}(\mathbf{i}, \mathbf{j} \in \mathbf{KCs}, \mathbf{s}, \mathbf{f}) = \sum_{\mathbf{j} \in \mathbf{common\ KCs}} (\beta_j + \gamma_j s_{i,j} + \rho_j f_{i,j}) + \sum_{\mathbf{j} \in \mathbf{rare\ KCs}} (\beta_d + \gamma_d s_{i,j} + \rho_d f_{i,j})$$

As discussed above, our data set has 27 common skills, and 37 rare skills. We trained our *merged-rare* model with 28 sets of parameters: 27 for the common benchmarks, and one default set that applied for every rare benchmark. No degenerate parameters of any type were observed. Validating the model against our testing dataset resulted in an AUC of 0.7836 and an RMSE of 0.4223 for the entire testing dataset, slightly better than the results of our baseline PFA model.

An alternative approach to handling rare benchmarks is to use the average parameters of the common skills for the rare skills. This model obtains an AUC of 0.7834 and an RMSE of 0.4223 for the entire test set, essentially the same results – though again, no degeneracy. This suggests that using a set of parameters that is calculated as an average across the trained learned parameters might be a reasonable option to handle entirely new or very rare skills.

Overall, then, combining rare skills in PFA can reduce degeneracy when data is limited.

4.3 Compensatory vs Conjunctive Skills in PFA

Another question is whether skills are conjunctive or compensatory. To succeed when solving an item with conjunctive skills, a student needs to know all the skills tagged to that item. For an item that has compensatory skills, knowing only a subset of the tagged skills could be enough.

PFA’s original formula is inherently compensatory. PFA can be adjusted to be conjunctive by multiplying together the skills rather than adding them (i.e. replacing the $\sum$ with $\prod$ in the m function). A third, in-between assumption is possible – that performance depends on each skill evenly. This *even-skill* model, uses averaging instead of $\sum$ or $\prod$ in the function. When fitting our data with the compensatory model we see an AUC of 0.7818 and a RMSE of 0.4245 on the testing dataset, whereas when using a conjunctive model the performance drops significantly, resulting in an AUC of 0.6725 and a RMSE of 0.4666. The performance results for the even-skill approach was slightly better than the compensatory approach, with an AUC of 0.7849 and an RMSE of 0.4171. This result contradicts past findings for BKT, where conjunctive models performed best in other data (i.e. Pardos et al, 2008).

5. Discussion and Conclusions

In this paper, we have discussed some of the challenges in real-world use of PFA. We focus on four potential areas of challenge: insufficient number of practices, degenerate parameters, rare benchmarks, and compensatory vs. conjunctive skill relationships. Overall, we find that degenerate parameters are an issue for PFA, particularly for benchmarks where there is limited initial data per student, and that degeneracy can be hidden by overall high performance (much like BKT – cf. Baker et al., 2008).

Other challenges may emerge in using PFA in real-world settings, depending on the design of the learning system. For one thing, the tagging of skills to items – the knowledge structure -- may have flaws, and there may be benefits to refitting the knowledge structure. Of course, the cost to arbitrarily refitting item to skill mappings is interpretability, and other challenges listed above – such as model degeneracy -- may also become more prominent with more complex mappings. These challenges may magnify if the skills have pre-requisite structure, but the order of the content does not respect the pre-requisite structure. In that case, a skill may be encountered both before and after a student has mastered its prerequisites, leading to sudden spikes in learning which PFA may be unable to capture, in contrast to BKT (which assumes sudden shifts in knowledge) or DKT-family algorithms (which can capture arbitrarily complex relationships between items). Another potential opportunity is seen in cases (like this data set) where tags are at the level of standards or benchmarks rather than fine-grained skills

– in these cases, making the tagging more fine-grained may increase the precision of estimation.

Even if items are tagged well to skills, items may have different levels of difficulty due to issues such as cognitive load or calculation time. PFA has a natural way of handling this issue – shifting the β parameter to the item level, or adding multiple β parameters (one at the skill level, one at the item level). Differences in item guess and slip may also play a role. Unlike BKT, which incorporates guess and slip probabilities for each skill, the original version of PFA does not take that into account (see MacLellan, Liu, & Koedinger, 2015 for a variant that does). While there is a 50% chance for a student to answer a true/false question correctly by guessing, the chances are smaller for a multiple choice question, and far smaller for a fill in the blank question, suggesting that the item type could be useful to model.

Overall, this article demonstrates that there are several considerations that must be taken into account in using PFA in the real-world, as previously demonstrated for other learning algorithms. However, for our specific data set, PFA's limitations seem very feasible to address with only minor adjustments. Given PFA's high interpretability, predictable behavior, and competitive performance in terms of AUC ROC/RMSE, PFA is a very reasonable choice for real-world student modeling.

References

Baker, R.S.J.d., Corbett, A.T., Aleven, V. (2008) More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (pp. 406-415). Montreal, Canada: Springer.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, *4*(4), 253-278.

Gervet, T., Koedinger, K., Schneider, J., & Mitchell, T. (2020). When is Deep Learning the Best Approach to Knowledge Tracing?. *Journal of Educational Data Mining, 12* (3), 31-54.

Gong, Y., & Beck, J. E. (2011). Looking beyond transfer models: finding other sources of power for student models. *Proc. of the Int.l Conf. on User Modeling, Adaptation, and Personalization* (pp. 135-146).

Gong, Y., Beck, J. E., & Heffernan, N. T. (2010). Comparing knowledge tracing and performance factor analysis by using multiple model fitting procedures. *Proc. of the International conference on intelligent tutoring systems* (pp. 35-44). Pittsburgh, PA, USA: Springer.

Lee, S., Choi, Y., Park, J., Kim, B., Shin, J. (2021) Consistency and monotonicity regularization for neural knowledge tracing, Preprint. Retrieved 3/27/2021 from https://openreview.net/forum?id=4P35MfnBQIY.

MacLellan, C. J., Liu, R., & Koedinger, K. R. (2015). Accounting for Slipping and Other False Negatives in Logistic Models of Student Learning. *Proc. of the Int'l Conf on Educational Data Mining*. Madrid, Spain.

Pardos, Z. A., Beck, J., Ruiz, C., Heffernan, N. T. (2008) The Composition Effect: Conjunctive or Compensatory? An Analysis of Multi-Skill Math Questions in ITS. *Proc. of the Int Conf. on Educational Data Mining*.

Pardos, Z., & Heffernan, N. (2010). Navigating the parameter space of Bayesian Knowledge Tracing models: Visualizations of the convergence of the Expectation Maximization algorithm. In *Proceedings of the International Conference on Educational Data Mining* (pp. 161-170). Pittsburgh, PA, USA.

Pavlik, P.I., Cen, H., Koedinger, K.R. (2009) Performance Factors Analysis – A New Alternative to Knowledge Tracing. *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 531-538).

Pelánek, R., Rihák, J., & Papoušek, J. (2016). Impact of data collection on interpretation and evaluation of student models. *Proc. of the 6th int. conf. on learning analytics & knowledge* (pp. 40-47).

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. *In Proc. of the 28th Int. Conf. on Neural Information Processing Systems* (pp. 505-513). Montreal, Canada: ACM Press.

Sales, A. C., & Pane, J. F. (2019). The role of mastery learning in an intelligent tutoring system: Principal stratification on a latent variable. *The Annals of Applied Statistics*, *13* (1), 420-443.

Scruggs, R., Baker, R.S., McLaren, B.M. (2020) Extending Deep Knowledge Tracing: Inferring Interpretable Knowledge and Predicting Post System Performance. *Proc. of the 28th Int. Conf. on Computers in Ed*.

Slater, S., Baker, R.S. (2018) Degree of Error in Bayesian Knowledge Tracing Estimates From Differences in Sample Sizes. *Behaviormetrika*, *45* (2), 475-493.

Van de Sande, B. (2013). Properties of the Bayesian Knowledge Tracing Model. *Journal of EDM*, *5* (2), 1-10.

Yeung, C. K. (2018). Improving deep knowledge tracing with prediction-consistent regularization. Doctoral dissertation, Hong Kong University of Science and Technology.

Yudelson, M. (2019). Elo, I Love You Won't You Tell Me Your K. *European Conference on Technology Enhanced Learning* (pp. 213-223). Delft, The Netherlands: Springer.

Zhang, J., Shi, X., King, I., & Yeung, D. Y. (2017). Dynamic key-value memory networks for knowledge tracing. *Proc. of the 26th int. conf. on World Wide Web* (pp. 765-774). Perth, Australia: ACM Press.

Does Large Dataset Matter? An Evaluation on the Interpreting Method for Knowledge Tracing

Yu LU[a, b], **Deliang WANG**[a*], **Penghe CHEN**[b] **& Qinggang MENG**[b]
[a]*School of Educational Techonology, Beijing Normal University, China*
[b]*Advanced Innovation Center for Future Education, Beijing Normal University, China*
*wangdeliang97@mail.bnu.edu.cn

Abstract: Deep learning has become a competitive method to build knowledge tracing (KT) models. Deep learning based knowledge tracing (DLKT) models adopt deep neural network but lack interpretability. The researchers have started working on interpreting the DLKT models by leveraging on methods in explainable artificial intelligence (xAI). However, the previous study was conducted on a relatively small dataset without comprehensive analysis. In this work, we perform the similar interpreting method on the largest public dataset and conduct the comprehensive experiments to fully evaluate its feasibility and effectiveness. The experiment results reveal that the interpreting method is feasible on the large-scale dataset, but its effectiveness declines with the larger size of learners and longer sequences of learner exercise.

Keywords: Knowledge tracing, deep learning, explainable artificial intelligence

1. Introduction

Knowledge tracing (KT) attempts to model learners' dynamic knowledge states on the skill level and predict their performance on the following exercises. With strong capacity to learn the inherent relationships from exercise data, deep learning has been adopted to build KT models. However, deep learning based knowledge tracing (DLKT) models have an untransparent decision process impeding their deployment. By leveraging on a technique called layer-wise relevance propagation (LRP) (Bach et al., 2015), we explored interpreting the DLKT model on a small dataset (Lu et al., 2020). It is still an open question whether the post-hoc interpreting method is feasible on large datasets.

In this work, we adopt the LRP method on one of the largest datasets, called EdNet (Choi et al., 2020). Specially, we clarify the technique of the LRP method in section 3, and perform the experiments to evaluate the feasibility and effectiveness of the method in section 4. The results reveal the LRP method is feasible on the large dataset, but its effectiveness declines with the larger size of learners and longer exercise sequence. By demonstrating the effectiveness issues of the current interpreting method, this work would be a solid step to build a fully transparent DLKT models.

2. Related Work

2.1 Knowledge Tracing

Bayesian knowledge tracing (BKT) (Corbett & Anderson, 1995) can be regarded as the most prominent KT model, adopting the hidden Markov model (HMM) to estimate learner's mastery state on individual skill. Subsequent studies consider more factors to improve BKT, e.g., knowledge prior (Chen et al., 2017). Besides, logistic regression models have been deployed to build KT models. Recently, deep learning was introduced into KT domain. Deep knowledge tracing (DKT) (Piech et al., 2015) was the pioneer work. Then, the dynamic key-value memory network (DKVMN) (Zhang et al., 2017) and its variants (Chaudhry et al., 2018) were adopted to improve model performance. The attention network (Su et al., 2018) had been adopted to better represent question semantics. Besides, other information, e.g., prerequisite information (Chen et al., 2018), was utilized to design new DLKT models.

2.2 *Explainable AI*

The intransparent decision process of deep learning models is often hard to understand for human. To tackle this issue, researchers have proposed many explainable AI methods to interpret models' outputs and their inner working mechanism. The interpretability can be classified as ante-hoc and post-hoc: the ante-hoc interpretability focuses on training simple-structured machine learning models (Melis & Jaakkola, 2018), e.g., linear regression. The post-hoc interpretability focuses on interpreting the trained models. Among the post-hoc interpretability, the local methods such as backward propagation (Zeiler & Fergus, 2014) mainly aim to clarify the importance of the input features to model's predictions. In this work, we adopt a backpropagation method, namely LRP method, to interpret the DLKT models.

3. Building and Interpreting DLKT Models

3.1 *DLKT Models on EdNet and ASSISTments*

We adopt the long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) to build the DLKT models in this work. Figure 1 illustrates the basic architecture of the DLKT model on EdNet. The interaction between learner and question can be denoted as the question-answer pair $x_t = \{(q_t, a_t) | \mathrm{t} = 1, \dots, \mathrm{N}\}$, where q_t is the representation of question information, $a_t \in \{0,1\}$ is the binary representation of correct or false answer, and N > 0 is the sequence length. The LSTM accordingly maps the input sequence vectors $\{\dots x_{t-1}, x_t, x_{t+1} \dots\}$ to the output vector $\{\dots y_{t-1}, y_t, y_{t+1} \dots\}$. Given most of the individual questions in EdNet covering multiple skills, an additional layer is adopted, which simply sets the average probabilities of all the skills covered by the next question as the final prediction z_t as below:

$$z_t = \frac{y_t \cdot q_{t+1}}{\mathrm{m}}, \tag{1}$$

where the dot product operation is performed between y_t and q_{t+1}, and *m* is the number of skills in next question. For ASSISTments dataset (Feng et al., 2009), this additional layer is not necessary.

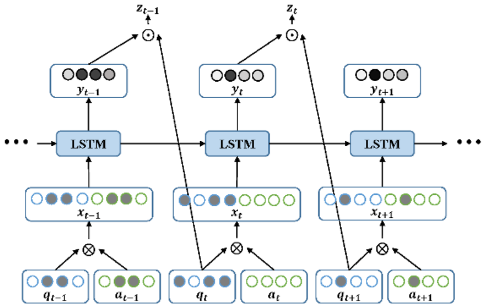


Figure 1. The Architecture of a RNN-based DLKT Model.

3.2 *Interpreting Method*

The LRP method interprets the DLKT models by analyzing the contribution of the individual input to the model's final prediction. Given a prediction made by the DLKT model, the LRP method would first sets the model's prediction value as the output layer neuron's relevance, and then backpropagate the relevance from the output layer to the input layer. During the backpropagating process, it needs to handle two different connections in the intermediate layers, namely *weighted linear connection* and *multiplicative connection*. The *weighted linear connection* can be written in a general form:

$$a_j^{(l+1)} = \sum_i w_{ij} \ a_i^{(l)} + b_j \tag{2}$$

where $a_j^{(l+1)}$ is the information the neuron *j* in layer *l+1* receives from the forward direction, w_{ij} and b_j are the weight and bias term. Given the relevance the neuron *i* in layer *l* receives from the neuron *j* in the layer *l+1* is $R_{i \leftarrow j}^{(l)}$, we have

$$R_{i\leftarrow j}^{(l)} = \frac{w_{ij}a_i^{(l)} + \frac{sign\left(a_j^{(l+1)}\right)\varepsilon + b_j}{N}\delta}{a_j^{(l+1)} + sign\left(a_j^{(l+1)}\right)\varepsilon} * R_j^{(l+1)} \tag{3}$$

where N is the number of neurons in layer l, and the item $sign\left(a_j^{(l+1)}\right) * \varepsilon$ prevents $R_{i\leftarrow j}^{(l)}$ becoming unbounded with $sign\left(a_j^{(l+1)}\right)$ as 1 or -1 and ε as a small positive value. We set δ as 0 to conserve relevance for the lower-level neurons. $R_j^{(l+1)}$ is the total relevance of neuron j in the layer $l+1$. For multiplicative connections, we define the neuron whose output ranging between 0 to 1 as "gate" neuron, and the remaining one as the "source" neuron. The multiplicative connection can be written as:

$$a_j^{(l+1)} = a_g^{(l)} \odot a_c^{(l)} \tag{4}$$

where $a_g^{(l)}$ and $a_c^{(l)}$ respectively are the message the "gate" neuron g and the "source" neuron s receive from layer l. During the forward propagating process, the "gate" neuron decides how much of the information should be retained in the upper-layer neurons and contributed to the model's decision (Arras et al., 2017). We set its relevance $R_{g\leftarrow j}^{(l)}$ as zero and give the full credit $R_j^{(l+1)}$ to the "source" gate.

4. Evaluation

4.1 Datasets and DLKT Models

We choose ASSISTment2009 and EdNet as the two datasets for the experiments. Table 1 summarizes the statistics of the preprocessed datasets. The built DLKT models adopt the LSTM network and RMSprop optimization for model training, with the iteration number and learning rate as 500 and 0.01. We set the hidden dimensionality, mini-batch size and the dropout rate to 200, 100 and 0.5 respectively. For both datasets, we utilize 64% data for training, 16% data for validating and the remaining ones for testing. After five-fold cross-validation, overall prediction accuracy (ACC) and AUC achieve 0.70 and 0.73 for the DLKT model on ASSISTment2009, and achieve 0.68 and 0.66 on EdNet.

Table 1. *Statistics of the Preprocessed Two Datasets ASSISTment2009 and EdNet*

Dataset	Learners	Skills	Questions	Interactions
ASSISTment 2009	3,091	110	16,850	320,582
EdNet	442,030	188	12,372	93,359,825

4.2 Feasibility Evaluation

4.2.1 Consistency Experiment

Given the calculated relevance for each question-answer pair, we investigate whether the sign of the relevance is consistent with the correctness of the answer. We define correctly-answered questions with positive relevance or falsely-answered questions with negative relevance as *consistent questions* and define the percentage of the *consistent question* in each sequence as *consistent rate*. A high consistent rate reflects that the LRP method could properly differentiate the correctly-answered and falsely-answered questions. Specifically, we utilize 7,143 and 1,187,377 sequences with a length of 15 in the two datasets as the test data. For each sequence, the first 14 question-answer pairs are the input, and the last pair is to validate the model's prediction. We obtain 4,972 correctly-predicted sequences in ASSISTment2009 and 799,857 correctly-predicted sequences in EdNet.

Figure 2 gives the histogram of the consistent rate on the two datasets. Nearly 80% sequences achieve a high consistent rate (i.e., 90% or above) in ASSISTment2009, while only around 50% sequences achieve a high consistent rate (i.e., 90% or above) in EdNet. Both distributions clearly show that the majority of sequences in both datasets receive 70% consistent rate or above, which demonstrate the sign of the calculated relevance values on both the small and large datasets.

4.2.2 Deletion Experiment

We further quantitatively investigate the relevance for both datasets by performing the deletion experiment. Specifically, for each correctly-predicted sequence, we delete the question-answer pair in a decreasing order of their relevance values for positive predictions or in an increasing order for negative predictions, and then record the predictions accuracy after each deletion. We also delete the question-answer pairs at random for comparison. Figure 3 illustrates the results on ASSISTment2009 and EdNet. For both datasets, all the accuracy lines drop down from 1.0 with an increasing number of the question-answer pair deletions, but the LRP lines drop much faster than the random lines.

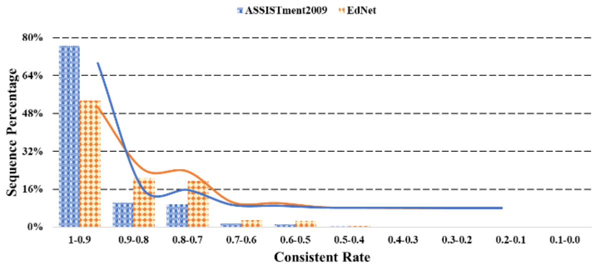
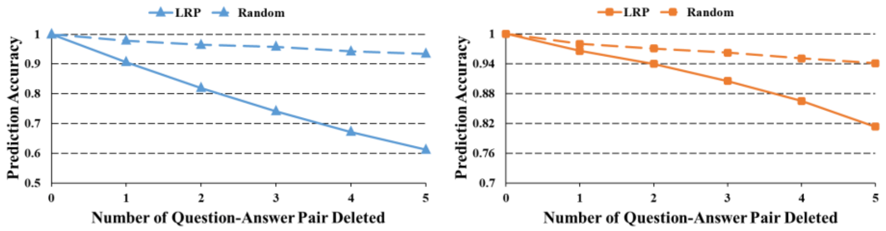
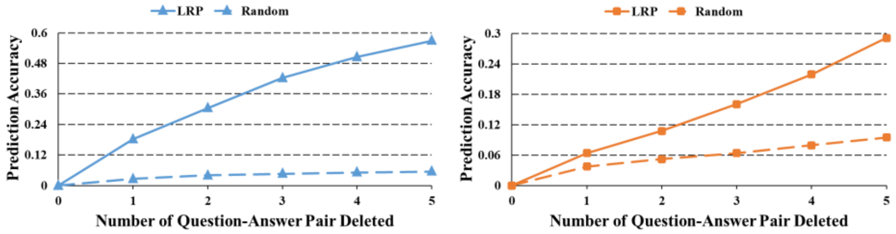


Figure 2. Histogram of the Consistent Rate on ASSISTment2009 and EdNet.



(a) ASSISTment2009 (b) Ednet

Figure 3. Accuracy Changes of Correctly-Predicted Sequences on the Two Datasets.



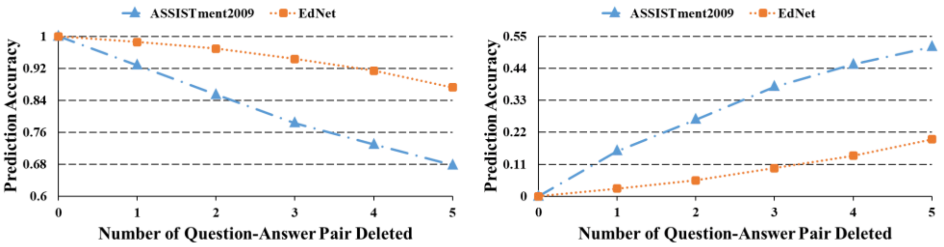
(a) ASSISTment2009 (b) Ednet

Figure 4. Accuracy Changes of Falsely-Predicted Sequences on the Two Datasets.

For each falsely-predicted sequence, we conduct similar experiments. Figure 4 shows that the accuracy lines rise from 0.0 with an increasing number of deletions for both datasets, but the LRP lines rise much faster than the random lines. All the deletion experiment results illustrate that the quantity of the relevance computed by the LRP method is possible to infer the question-level contribution to the final prediction result, and the LRP method is feasible on the large dataset EdNet.

4.3 Effectiveness Evaluation

To evaluate the effectiveness of the LRP method, we compare the accuracy changes between ASSISTment2009 and EdNet directly, deducting the random deletion effect. Figure 5 shows that the accuracy changes on ASSISTment2009 are much larger than on EdNet, which indicates the LRP method is less effective on the large dataset. This might be due to the larger number of learners. Another feature of the EdNet is the length of its sequences, which are larger than the ones in ASSISTment2009. We design the experiments on EdNet to evaluate whether the length of sequences (i.e., the number of question-answer pairs in one sequence) affect the effectiveness of the LRP method. Specifically, we divide the test data in EdNet into the sequences at a length of 15, 50, 100 and 200, and conduct the consistency and deletion experiments Table 2 summarizes sequence number at different lengths.



(a) Correctly-Predicted Sequences (b) Falsely-Predicted Sequences

Figure 5. Comparison of the Accuracy Changes between ASSISTment2009 and EdNet.

Table 2. *Number of Sequences at a Length of 15, 50, 100 and 200 in EdNet*

		Len15	Len50	Len100	Len200
Correctly Predicted	Positive Prediction	751,737	218,173	102,603	47,110
	Negative Prediction	48,120	11,698	5,384	2,467
Falsely Predicted	Positive Prediction	354,901	96,491	45,097	20,289
	Negative Prediction	32,619	8,540	3,908	1,745
Total		1,187,377	334,902	156,992	71,611

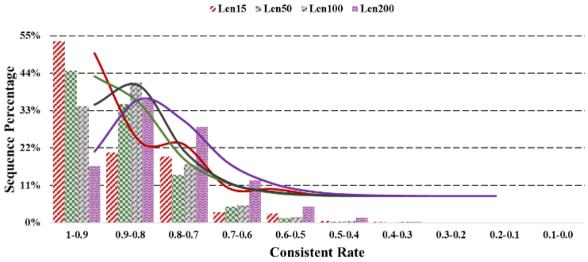
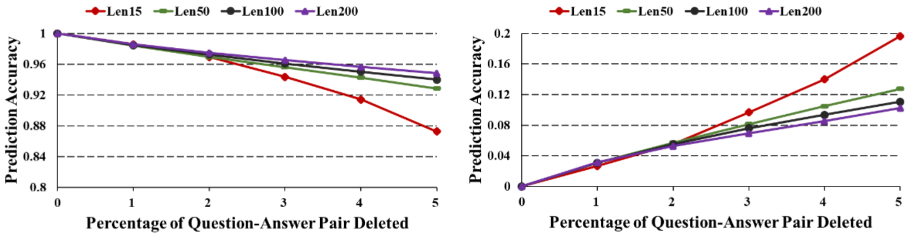


Figure 6. Histogram of the Consistent Rate at Different Lengths of the Sequences.



(a) Correctly-Predicted Sequences (b) Falsely-Predicted Sequences

Figure 7. Comparison of the Accuracy Changes at Different Lengths of the Sequences.

Figure 6 gives the histogram of the consistent rate at different lengths. Less than 5% sequences at different lengths have a consistent rate below 0.6, showing the feasibility of the LRP method. Different lengths exhibit distinct distributions: for the shorter ones, e.g., length of 15 and 50, the highest sequence percentage bars appear in the consistent rate range 1-0.9, and then they sharply drop down with the decreasing consistent rate. For the longer sequences, e.g., length of 100 and 200, the highest sequence percentage bars appear in the consistent rate range 0.9-0.8, and then drop down smoothly. In other words, the distributions tend to display their non-monotonic and long-tail patterns, which indicates that the long sequences might affect the sign of the relevance calculated by the interpreting method.

Figure 7 presents the deletion experiment results at different lengths. For the correctly-predicted sequences, Figure 7(a) shows all the accuracy lines drop down with an increasing number of deletions. However, the lines of shorter sequences (e.g., length of 15) drop much faster than the longer ones (e.g., length of 200). For the falsely-predicted sequences, in Figure 7(b) all the accuracy lines rise up with an increasing number of deletions. However, the lines of shorter sequences (e.g., length of 15) rise much faster than the longer ones (e.g., length of 200). Both experiment results indicate that the relevance for longer sequences are more difficult to reflect the question-level contributions to the prediction. It is more difficult for the LRP method to capture important question-answer pairs from longer sequences.

5. Conclusion

In this work, we first build the RNN-based DLKT models on both ASSISTment2009 and EdNet, and then perform the LRP methods on both the small-scale and large-scale models. Both the consistency and deletion experiments validate the feasibility of the interpreting method on the large dataset EdNet. However, the current interpreting method performs less effective on EdNet, which might be mainly due to its bigger size of learners and longer sequence of learner exercise. On a broader canvas, this work validates the basic interpreting method for explaining the DLKT model's predictions, but suggests the new studies to improve the current interpreting methods due to the large-scale educational datasets.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No. 62077006, 61807003), and the Fundamental Research Funds for the Central Universities.

References

Arras, L., Montavon, G., Müller, K. R., & Samek, W. (2017). Explaining recurrent neural network predictions in sentiment analysis. *EMNLP 2017,*page 159.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Muller, K., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *Plos One*, 10(7):0130140.

Chaudhry, R., Singh, H., Dogga, P., & Saini, S. K. (2018). Modeling hint-taking behavior and knowledge state of students with multi-task learning. In *Proceedings of Educational Data Mining*.

Chen, P., Lu, Y., Zheng, V. W., & Pian, Y. (2018). Prerequisite-driven deep knowledge tracing. In *2018 IEEE International Conference on Data Mining (ICDM),* pages 39–48. IEEE.

Chen, Y., Liu, Q., Huang, Z., Wu, L., Chen, E., Wu, R., Su, Y.,& Hu, G. (2017). Tracking knowledge proficiency of students with educational priors. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management,* pages 989–998. ACM.

Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., Baek, J., Bae, C., Kim, B., & Heo, J. (2020). Ednet: A large-scale hierarchical dataset in education. In *International Conference on Artificial Intelligence in Education*, pages 69–73. Springer.

Corbett, A. T. & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-adapted Interaction*, 4(4):253–278.

Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modelingand User-Adapted Interaction*, 19(3):243–266.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.

Lu, Y., Wang, D., Meng, Q., & Chen, P. (2020). Towards interpretable deep learning models for knowledge tracing. In *International Conference on Artificial Intelligence in Education*, pages 185–190. Springer.

Melis, D. A. & Jaakkola, T. S. (2018). Towards robust interpretability with self-explaining neural networks. In *Proceedings of Advances inNeural Information Processing Systems(NIPS),* pages 7786–7795.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In *Advances in NeuralInformation Processing Systems*, pages 505–513.

Su, Y., Liu, Q., Liu, Q., Huang, Z., Yin, Y., Chen, E., Ding, C., Wei, S., & Hu, G. (2018). Exercise-enhanced sequential modeling for student performance prediction. In *32nd AAAI Conference on Artificial Intelligence*.

Zeiler, M. D. & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proceedings of European Conference on Computer Vision*, pages 818–833.

Zhang, J., Shi, X., King, I., & Yeung, D. Y. (2017). Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th International Conference on World Wide Web*, pages 765–774.

Using Qualitative Data from Targeted Interviews to Inform Rapid AIED Development

Jaclyn OCUMPAUGH[a*], Stephen HUTT[a], Juliana Ma. Alexandra L. ANDRES[a], Ryan S. BAKER[a], Gautam BISWAS[b], Nigel BOSCH[c], Luc PAQUETTE[c] & Anabil MUNSHI[b]
[a]*University of Pennsylvania, USA*
[b]*Vanderbilt University, USA*
[c]*University of Illinois at Urbana Champaign, USA*
*ojaclyn@upenn.edu

Abstract: This paper examines how interviews with students—at critical moments of the learning process—may be leveraged to improve the design of educational software. Specifically, we discuss iterative work to improve the design of a pedagogical agent in the Betty's Brain learning environment, Mr. Davis. Students interacted with the pedagogical agent in Betty's Brain during two separate studies, two months apart. During study one, qualitative interviews were prompted by student actions within the system and theoretically aligned sequences of educationally relevant affective states (as detected by previously validated models). Facilitaed by an app called the Quick Red Fox (QRF), these *in situ* interviews were then used to identify ways to rapidly improve Mr. Davis' design, investigated in study two. Results indicate that changes designed to make Mr. Davis more empathetic correlate with improved learning outcomes. We also discuss the potential for rapidly collected qualitative data in future developments.

Keywords: Affective computing, learning by teaching, data-driven design, pedagogical agents

1. Introduction

Computer-based learning systems that seek to imitate human tutoring are faced with considerable challenges related to the design of their virtual pedagogical agents. Human tutors can dynamically adjust their interactions with students, accommodating differences in age, personality, cultural expectations, as well as moment-to-moment differences in affect or understanding. In contrast, virtual pedagogical agents are far less flexible, typically relying on a predefined set of tutorial actions or phrases (Veletsianos & Russell, 2014). Since it is not always possible to anticipate how students will interact with an agent (Kenkre & Murthy 2016), researchers need a way to quickly identify potential changes in student behavior and suggest improvements.

Pedagogical agents offer great potential for improving learning and supporting learners. Early work has shown that virtual agents can support increases in performance (Dumdumaya et al. 2017), provide beneficial social interactions (Doering, Veletsianos, & Yerasimou 2008; Kim & Wei 2011), and foster motivation and engagement (Kim & Wei 2011; Lusk & Atkinson 2007). However, some aspects of what makes a good interaction with a virtual agent is still an open question. Research has sought to calibrate pedagogical agents based on politeness, with many students responding well to more polite approaches (Graesser 2011; Person et al. 1995; Tynan 2005), but there have sometimes been conflicting results. Some studies have found that learning decreases when students perceive their tutors as irritating or offensive (De Angeli & Brahnam 2008). However, this pattern is not universal; other work has also shown that in some contexts, students respond well to so-called "rude tutors" who are designed to offer sarcastic responses even to struggling students (e.g., Graesser, 2011; Ogan, Finkelstein, Mayfield, et al. 2012).

Conflicting responses to politeness strategies are perhaps not surprising. Given the wide range of the functions of politeness (and impoliteness) that human tutors use (see review in Ogan, Finkelstein, Walker, et al. 2012), cultural, contextual, and developmental differences in student populations would likely moderate these findings (Savard & Mizoguchi 2019). Cultural differences are, unfortunately, more difficult to tease out from the current literature than developmental differences (Paquette et al.

2020), but given findings that show high and low-knowledge learners respond differently to politeness strategies (D'Mello & Graesser 2013), developmental differences seem a likely explanation.

Thus far, most of the research on politeness and empathy has taken place with older adolescents and adult learners (Ogan, Finkelstein, Walker, et al. 2012; Rodrigo et al. 2013; Wang et al. 2012) (D'Mello & Graesser 2013). Less is known about the politeness practices of primary learners (ages 5-11), where students might be more likely to expect more empathetic pedagogical strategies.

This paper explores younger (6th grade) students' perception of two different versions of Mr. Davis, the mentor agent in the Betty's Brain learning system (Biswas, Segedy, & Bunchongchit 2016). Specifically, during *in situ* interviews with students, which were facilitated by an app called Quick Red Fox (QRF), we discovered that students were interpreting certain phrases being used by Mr. Davis as deliberately offensive (which was not the original intent of the designer), and unlike previous research on rude tutorial moves (e.g., Graesser 2011; Ogan, Finkelstein, Walker, et al. 2012), students were not responding well to these perceived slights. Accordingly, we followed up with further questions about what would make Mr. Davis feel more helpful. We then modified Mr. Davis's responses to make Mr. Davis seem more supportive. Critically, the insights occurred during initial reflection on the interviews, prior to more formal transcription and coding. This allowed for changes to be made before students used the software again two months later (for a new learning topic). Following this second session, students were surveyed about their current perceptions of Mr. Davis as well as their perceptions of how difficult the task was. In this paper, we report on what those changes are and then compare the perception of the two versions of Mr. Davis with learning gains as experienced by 99 middle schoolers across two studies.

2. Betty's Brain

Betty's Brain uses a learning-by-teaching model (Biswas et al. 2004), where students must teach a virtual agent named Betty by creating a causal map of a scientific process (e.g., climate change). Betty demonstrates her "learning" by taking quizzes that are graded by a pedagogical agent, Mr. Davis. As students construct Betty's map, they must navigate various learning resources, including hypermedia resources and a teaching manual that explains how to represent causal reasoning. The system is open-ended, with students choosing when to consult hypermedia resources, when to devote time expanding their causal map, and when to test it via the quizzes (Biswas et al. 2016).

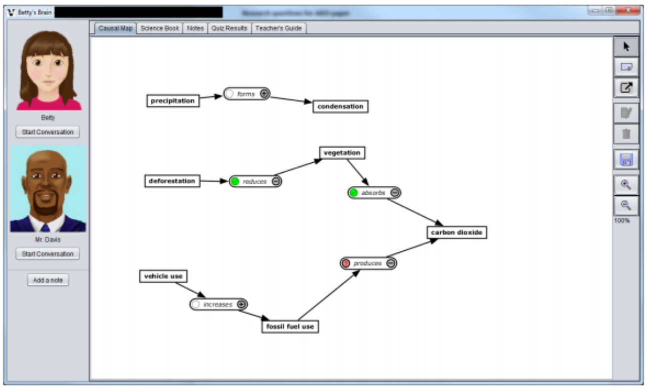


Figure 1. A Partial Casual Map in Betty's Brain.

Each time the students test their causal map (by testing Betty), Mr. Davis interacts with the students, giving them high-level suggestions on how they may improve. However, students can interact with him at other times in the learning process, asking him questions or seeking help/tips regarding causal maps. Mr. Davis is described to the students as an experienced teacher and is designed to present a helpful and mentoring role as the students teach Betty.

3. Methods

3.1 Data

Data were collected over two studies with 99 6th-graders who used Betty's Brain as part of their normal science instruction in an urban public school in the United States of America. In Study 1 (December 2018), students used Betty's Brain to learn about climate change. During the study, short interviews (usually 1-2 minutes) were repeatedly conducted, with researchers asking students to briefly pause their activity to talk about their work (more details in section 3.2). In Study 2 (February 2019), the same group of students interacted with Betty's Brain again and participated in a similar interview process as they learned and created a causal model for a new topic: thermoregulation.

Learning gains were assessed during both studies using pre- and post-test measures. For each study, the pre-test about the content (i.e., climate change or thermoregulation) knowledge was administered before students began working with the system. An identical post-test was administered at the end of the study. In this work, we characterize learning gains as post-test score – pre-test score.

3.2 Interviews

In both studies, interviews were dynamically prompted by real-time monitoring of affective and behavioral sequences (detected by previously validated models - see Jiang et al. 2018) that were being studied as part of a larger project on student affect and self-regulated learning (Bosch et al. 2021). In particular, key shifts in affect, such as *confusion → delight, confusion → frustration, frustration → engaged*, as well as *sustained* delight, were used to trigger interviews. Other interviews were prompted by specific behavioral sequences, including sequences that identified how effective the students' behaviors were. That is, actions like *editing the causal map* (as opposed to *opening a reading passage*) were further categorized by whether those edits were correct or incorrect. Thus, interviews were taking place during times that had been strategically identified to represent shifts in emotional states and key moments in the learning process.

Interviewers were directed to students experiencing one of these prompts by an Android-based app called Quick Red Fox (QRF), which collected metadata (deidentified student IDs and timestamps) while recording both the behavioral or affective sequence that triggered the interview and the interview itself. The interviewer then engaged the students with open-ended questions like "how are things going so far?" and "what strategies are you using to solve this problem?" Students were sometimes also asked about their intrinsic interest in science (e.g., "What is your favorite class?"), but in general, the interviewer let students guide the topics of interviews.

Though interviews were eventually manually transcribed and qualitatively analyzed (Bosch et al. 2021), this was not completed until after both rounds of data collection. In the interim, interviewers summarized their perceptions about trends and issues to the broader research and development team.

In the second study (after changes described in the next section were implemented), interviews were triggered and conducted using the same methods, but interviewers added explicit questions about the changes to the design of Mr. Davis.

3.3 Changes to Mr. Davis

During the Study 1 interviews, two scripts employed by Mr. Davis repeatedly emerged as particularly upsetting to the students, both of which involved his text beginning with the word "Hmph." Students consistently reported that they found Mr. Davis's interactions abrupt, frustrating, and, in some cases, distracting. Interviewers, therefore, asked students for suggestions on how to rewrite the scripts, including suggesting simply changing "Hmph" to "Hmm," which students found contemplative and less confrontational. Students also suggested that having Mr. Davis provide hints or encouragement would further improve his design. These design suggestions were reported to the development team, who made the appropriate changes to the conversations by incorporating ten new interactions and modifying two existing interactions. Three examples are shown in in Table 1.

Table 1. *Breakdown of Modifications to Mr. Davis*

Type of Change	Example	N
Add new feedback with hints/guidance	"Hey, from the quiz results, it looks like Betty may have some incorrect links on her map. You can mark those links as could be wrong in your map. Do you want to know more about marking links as could be wrong?"	7
Add new feedback with encouragement	"Looks like you are doing a good job teaching correct causal links to Betty! Make sure that you check her progress from time to time by asking her to take a quiz."	3
Representational change for politeness	"Hmph" → "Hmm"	2

3.4 Surveys of Helpfulness and Difficulty

At the end of Study 2, students were asked to rate the helpfulness of Mr. Davis during each scenario (reflecting on both Study 1 and Study 2) using a single-item Likert measure from 1 (strongly dislike) to 5 (strongly like). Students were also asked to rate the difficulty of each scenario (also using a single item Likert measure).

4. Results

We examine the effectiveness of the changes to Mr. Davis by comparing the Likert scales on the perceptions about him between the two scenarios. We then explore the relationship between these results in terms of learning and student perceptions of difficulty.

4.1 Perceptions of Mr. Davis

We first considered if the changes to Mr. Davis improved student perception of him. A paired samples Wilcoxon signed-ranks test showed a statistically significant improvement in how helpful Mr. Davis was perceived as being across the two scenarios (Study 1: *M* = 2.10, *SD* = 1.04, Study 2: *M* = 2.83, *SD* = 1.37; *Z*=*.43*, *p* < 0.001). This aligns with the qualitative reports given during student interviews in the second study. A histogram of student responses regarding Mr. Davis is shown in Figure 2.

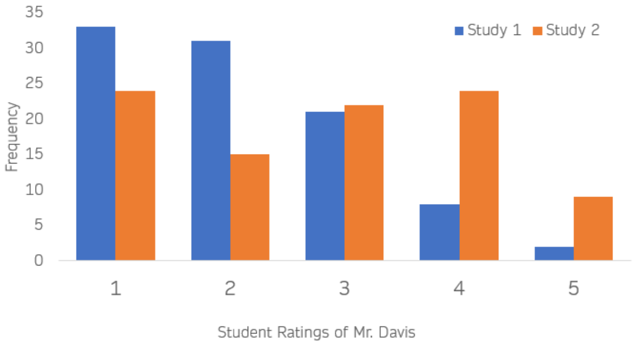


Figure 2. Bar Chart of Student Scores of Mr. Davis across both Studies (1 = Strong Dislike, 5 = Strong Like).

4.2 Mr. Davis and Learning Gains

We next examined how students' perceptions of Mr. Davis were linked to their learning gains within Betty's Brain. For each student, we calculated the difference between their two survey responses (Likert score for Study 2 minus Likert score for Study 1) with a positive score indicating that students felt Mr. Davis had improved. Through Spearman correlations, we found that improvement perception of Mr. Davis was positively correlated with learning gains in the second topic (rho = .37, p < 0.01).

To further examine if this result was an artifact of student self-efficacy (e.g. students being more likely to perceive Mr. Davis as helpful if they are doing well in the Betty's Brain), we examined the learning gains in the first scenario (prior to the changes to Mr. Davis) and observed no significant correlation to improvements in Mr. Davis (rho = .08, p = .46), reinforcing our conclusion that the changes to Mr. Davis had a positive impact on students' experiences. However, it should be noted that correlation is not causation, and other factors may have had an impact (see discussion).

4.3 Mr. Davis and Perceptions of Difficulty

Finally, we examined the difference in students' perceived difficulty of the two topics. We found that the change in students' perception of Mr. Davis was negatively correlated with change in perceived difficulty (*rho* = -.31, *p* < 0.01), indicating that students who found Mr. Davis more helpful the second time than the first time also found the second scenario easier than the previous scenario. There is insufficient evidence to be certain of causality, but this evidence is compatible with the idea that the students felt more supported during their second encounter with the system.

5. Discussion and Conclusions

Results show that the changes implemented between studies 1 and 2 led to more favorable interpretations of Mr. Davis. Specifically, we show a considerable increase in the perception of Mr. Davis as being helpful. Students' improved perceptions of Mr. Davis were also correlated with an improvement in learning gains and a reduction in perceived difficulty.

A limitation of this work is that we have not considered how additional external factors such as student familiarity with the system and topic familiarity may have impacted these perceptions due to scope. While our survey measures consider Mr. Davis in isolation, it is likely that more general perceptions about the system also have an impact here. More research is needed to determine how students' perceptions of Mr. Davis interact with other parts of the learning process (such as broader self-regulation), as is more research on what factors impact whether students prefer more polite and supportive pedagogical agents. However, this study demonstrates how even minor changes at the pragmatics level of conversation (i.e., changes that effect the perception of politeness and intent) may impact students' interactions with a learning system, suggesting that researchers should explore how politeness may interact with other parts of the design of computer-based learning environments.

This study also demonstrates the importance of exploring new methodologies for collecting feedback on student-tutor interactions. In particular, it suggests that rapidly collected qualitative data, such as those collected in the interviews prompted by the QRF app, might be useful for rapid design iteration and improvement in real-world settings. This is especially true in situations when implementing in a new context (e.g., with students in a new age group, or implementing in a new language). We often rely on quantitative data or formally coded interviews to drive design work, but this is often costly and time-consuming, meaning that relatively minor changes to educational software can take considerable time. This work presents a useful substitute for laboratory-based usability studies when those are impractical to conduct or may change students' attitudes towards the learning experience. While it sometimes difficult to anticipate what tutorial moves students will respond well to and exactly how to write tutorial dialogue, it can be very easy for a human interviewer to identify when a student responds poorly to a tutor's responses, allowing for timely adjustments to correct course.

Acknowledgments

This work was supported by NSF #DRL-1561567.

References

De Angeli, Antonella, & Sheryl Brahnam. 2008. "I Hate You! Disinhibition with Virtual Partners." *Interacting with Computers* 20(3):302–10.

Biswas, Gautam, Krittaya Leelawong, Kadira Belynne, Daniel Schwartz, & Joan Davis. 2004. "Incorporating Self Regulated Learning Techniques into Learning by Teaching Environments." in *The Twenty Sixth Annual Meeting of the Cognitive Science Society*.

Biswas, Gautam, James R. Segedy, & Kritya Bunchongchit. 2016. "From Design to Implementation to Practice a Learning by Teaching System: Betty's Brain." *International Journal of Artificial Intelligence in Education*.

Bosch, Nigel, Y. Zhang, Luc Paquette, Ryan Shaun Baker, Jaclyn Ocumpaugh, & Gautam Biswas. 2021. "Students' Verbalized Metacognition during Computerized Learning." P. 12 in *ACM SIGCHI: Computer-Human Interaction*. Association for Computing Machinery.

Cohen, J. 2013. *Statistical Power Analysis for the Behavioral Sciences*. Taylor & Francis.

D'Mello, Sidney K., & Art Graesser. 2013. "AutoTutor and Affective Autotutor: Learning by Talking with Cognitively and Emotionally Intelligent Computers That Talk Back." *ACM Transactions on Interactive Intelligent Systems* 2(4).

Doering, Aaron, George Veletsianos, & Theano Yerasimou. 2008. "Conversational Agents and Their Longitudinal Affordances on Communication and Interaction." *Journal of Interactive Learning Research* 19(2):251–70.

Dumdumaya, Cristina E., Michelle P. Banawan, Ma Rodrigo, T. Mercedes, Amy Ogan, Evelyn Yarzebinski, & Noboru Matsuda. 2017. "Investigating the Effects of Cognitive and Metacognitive Scaffolding on Learners Using a Learning by Teaching Environment." in *International Conference on Computers in Education (ICCE)*.

Graesser, Arthur C. 2011. "Learning, Thinking, and Emoting With Discourse Technologies." *American Psychologist* 66(8):746.

Jiang, Yang, Nigel Bosch, Ryan S. Baker, Luc Paquette, Jaclyn Ocumpaugh, Juliana Ma Alexandra L. Andres, Allison L. Moore, & Gautam Biswas. 2018. "Expert Feature-Engineering vs. Deep Neural Networks: Which Is Better for Sensor-Free Affect Detection?" Pp. 198–211 in *Artificial Intelligence in Education*.

Kenkre, A., & S. Murthy. 2016. "Students Learning Paths in Developing Micro-Macro Thinking: Productive Actions for Exploration in MIC-O-MAP Learning Environment." in *International Conference on Computers in Education (ICCE)*.

Kim, Yanghee, & Quan Wei. 2011. "The Impact of Learner Attributes and Learner Choice in an Agent-Based Environment." *Computers & Education* 56(2):505–14.

Lusk, Mary Margaret, & Robert K. Atkinson. 2007. "Animated Pedagogical Agents: Does Their Degree of Embodiment Impact Learning from Static or Animated Worked Examples?" *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 21(6):747–64.

Ogan, Amy, Samantha Finkelstein, Elijah Mayfield, Claudia D'Adamo, Noboru Matsuda, & Justine Cassell. 2012. "'Oh, Dear Stacy!' Social Interaction, Elaboration, and Learning with Teachable Agents." Pp. 39–48 in *Conference on Human Factors in Computing Systems*.

Ogan, Amy, Samantha Finkelstein, Erin Walker, Ryan Carlson, & Justine Cassell. 2012. "Rudeness and Rapport: Insults and Learning Gains in Peer Tutoring." Pp. 11–21 in *International Conference on Intelligent Tutoring Systems*.

Paquette, Luc, Jaclyn Ocumpaugh, Ziyue Li, Alexandra Andres, Ryan Baker, & others. 2020. "Who's Learning? Using Demographics in EDM Research." *JEDM| Journal of Educational Data Mining* 12(3):1–30.

Person, Natalie K., Roger J. Kreuz, Rolf A. Zwaan, & Arthur C. Graesser. 1995. "Pragmatics and Pedagogy: Conversational Rules and Politeness Strategies May Inhibit Effective Tutoring." *Cognition and Instruction* 13(2):161–69.

Rodrigo, Ma Mercedes T., R. Geli, Aaron Ong, G. Vitug, Rex Bringula, R. Basa, & N. Matsuda. 2013. "Exploring the Implications of Tutor Negativity towards a Synthetic Agent in a Learning-by-Teaching Environment." *Philippine Comput. J* 8:15–20.

Savard, Isabelle, & Riichiro Mizoguchi. 2019. "Context or Culture: What Is the Difference?" *Research and Practice in Technology Enhanced Learning* 14(1):23.

Tynan, Renee. 2005. "The Effects of Threat Sensitivity and Face Giving on Dyadic Psychological Safety and Upward Communication." *Journal of Applied Social Psychology* 35(2):223–47.

Veletsianos, George, & Gregory S. Russell. 2014. "Pedagogical Agents." Pp. 759–69 in *Handbook of research on educational communications and technology*. Springer.

Wang, William Yang, Samantha Finkelstein, Amy Ogan, Alan W. Black, & Justine Cassell. 2012. "'love Ya, Jerkface': Using Sparse Log-Linear Models to Build Positive (and Impolite) Relationships with Teens." Pp. 20–29 in *13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*.

Generating Student Progress Reports based on Keywords

Shumpei KOBASHI* & Tsunenori MINE
Kyushu University, Fukuoka, Japan
*kobashi.shumpei@m.ait.kyushu-u.ac.jp

Abstract: In this paper, we propose a method that automatically generates a student learning status report based on keywords given by instructors at cram schools to reduce their burden on writing the report. For selecting sentences to generate the report, we propose two methods: Seq2Seq-based and Information Retrieval (IR)-based methods. The Seq2Seq-based method uses a Seq2Seq model to generate sentences using keywords given by the instructors. The IR-based method uses OkapiBM25 to select sentences from those written by the instructors based on the keywords. We conducted extensive experiments to evaluate the two methods on a test set of 197,493 sentences. The experimental results show that the Seq2Seq method generates more suitable sentences as the report than the IR-based method. Adding the attention mechanism to the Seq2Seq method further improved the performance of the Seq2Seq method. Considering the above experimental results, we discussed the generation of the lecturer report by keywords.

Keywords: Natural language generation, keywords, education support,

1. Introduction

In many cram schools, instructors are required to write reports on students after each class. This report can be used for encouraging students to study and sometimes for handover between instructors. Meanwhile, writing reports increases the instructors' burden to think about and write the reports each time, which may lead to the deterioration of sentences in the reports. Consequently, it may increase mass-produced reports with less meaningful contents that are not suitable for student learning. Therefore, it is an important issue for both students and instructors to automatically or semi-automatically generate meaningful reports.

In this study, we propose methods to automatically generate instructor reports based on keywords, where we assume that the instructors can select appropriate keywords considering the status of student learning. We propose two methods: a sequence to sequence (Seq2Seq)-based and an Information Retrieval (IR)-based method. The Seq2Seq-based method uses a model called the Seq2Seq model which consists of an encoder and a decoder (Sutskever et al., 2014); the encoder transforms a sequence, a set of keywords in this study, to a latent vector z mapped on a latent space and the decoder receives z and decodes it to another sequence, the report sentence in this study. The IR-based method uses OkapiBM25 (Robertson et al., 1995) and selects sentences from those written by the instructors based on the keywords. In this study, we conduct extensive experiments on a test set of 197,493 sentences given by the instructors at actual cram schools to examine the performance of the two methods. The experimental results show that the Seq2Seq-based method generates more proper sentences as the reports than the IR-based method. Adding the attention mechanism to the Seq2Seq-based method further improved the performance of the Seq2Seq method.

The main contributions in this study are as follows: (1) we pointed out the necessity to automatically generate report sentences so as to reduce instructors' burden, (2) we proposed two methods: Seq2Seq-based and IR-based methods to generate the sentences, and (3) we conducted extensive experiments on the real dataset of report sentences written by instructors at cram schools to compare the performance of the two methods.

In what follows, Section 2 shows literature reviews and discusses the position of this study; Section 3 explains proposed models; Section 4 discusses experimental results. Finally, we conclude and discuss our future work.

2. Related Work

Research and development on systems to support instructors in educational institutions have been conducted for a long time. The duties of instructors include teaching students, preparing questions, and grading tests. Gutl et al. (Gutl et al., 2011) survey and present various types of research on the automatic creation of test items. The latest automatic generation of test questions has been shown to be comparable in quality to human-created questions, contributing to the elimination of the need for instructors to create test questions. The purpose of this is to reduce the time required to create question materials, i.e., to reduce the burden on the instructor. Liang et al. (Liang et al., 2018) proposed a neural network-based Automated Essay Scoring (ASE) model to automatically grade essays so as to reduce the manual workload of instructors and to provide rapid feedback on learning. Considering the effectiveness of textual feedback and its human burden, Lu et al. (Lu et. al., 2021) have implemented an ASE model that combines word-embedding and a deep learning model, and then developed a text-based automatic feedback system using the Constrained Metropolis-Hastings Sampling sentence paraphrase unsupervised learning. Also, Malik et al. (Malik et al., 2021) introduced a generative grading system that provides nuanced and interpretable feedback by modeling the student's response process and learning the student's reasoning process. This system showed promising results across multiple modalities and domains. The research described above aims to reduce the burden on instructors by fully automating their work. In other words, we can say that we are developing robot (AI) instructors. However, Parab (Parab, 2020) shows that human instructors are more comfortable for students than robot instructors, and that it is beneficial for the system to support human instructors in their daily work. Based on the above, this research aims to support the daily work of instructors, namely “report generation” by using language processing technologies. Therefore, the goal in this study is not to completely automate the lecturers’ work, but to reduce the workload of the lecturers while placing emphasis on the knowledge and insights of the human lecturers.

3. Proposed Methods to Generate Report Sentences based on Keywords

3.1 Proposed Methods

In this study, we propose two methods for generating sentences from keywords based on report sentences written by instructors at cram schools: one is the IR-based method, and the other is the Seq2Seq-based method. The IR-based method searches for past report sentences written by instructors using keywords, and selects the report with the highest rank. The Seq2Seq-based method learns a model to convert keywords to a report sentence based on report sentences written by instructors, and uses the model to generate a report sentence from keywords. Ideally, a comparison experiment between the two methods should be conducted using the following two sets of data: report sentences that instructors actually wrote and keywords that the instructors wanted to use to generate the report sentences. However, we can only use the data of the actual report sentences written by the instructors, which we call “original report data.” Therefore, assuming the target report sentences would include keywords given by instructors, we extracted the keywords corresponding to each report sentence from the original report data and also assumed that the extracted keywords can be regarded as the keywords given by the instructors. Using the two sets of data: the original report data and keywords extracted from the original report data, we invent a task if report sentences in the original report data can be generated from the keywords extracted from the original report data, and evaluate the performance of the two proposed methods.

3.2 Keyword Extraction from Report Sentence

In this study, we use 197,493 report sentences provided by actual cram schools, which were written by instructors for each class regarding the learning status of their students. The average number of words per sentence is 15.6, which is shorter than a typical sentence. Keywords are extracted from each report sentence using Term Frequency and Inverse Documentation Frequency (tf-idf) weights, which are commonly used in keyword extraction. The specific procedure of keyword extraction is described as follows:

1. We apply morphological analysis to each sentence and divide it into morphemes. We use "MeCab"[1] as the morphological analyzer.
2. To calculate the tf-idf weights of the morphological data, we use TfidfVectorizer in the feature_extraction module of scikit-learn[2]. The tf-idf weights are calculated for the data excluding auxiliary verbs, particles, conjunctions, and interjection because these parts of speech are rarely used as keywords.
3. Words with tf-idf weight values above the average are adopted as keywords of the report sentences. The average number of keywords per sentence is 5.14. This means that about 1/3 of the words in the original sentence were extracted as keywords.

3.3 *Overview of the Comparison Experiment*

To evaluate the performance of the two models built by the two proposed methods on generating or selecting sentences, we conduct 10-fold cross-validation whose overview is shown in *Figure 1*. The training and test data are divided into 9 to 1 and the model is built from the training data, and the test data is used to evaluate the model built in the training phase, and the performance of the model is evaluated using the evaluation metrics described below. This cycle is repeated 10 times and we take average of the results.

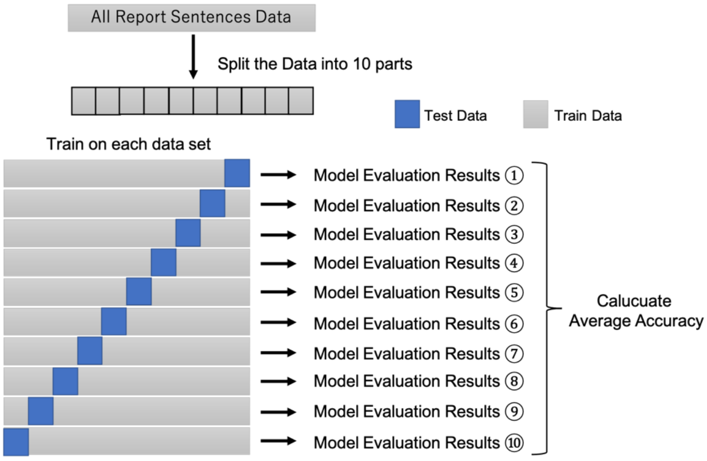

All Report Sentences Data
Split the Data into 10 parts
Test Data
Train Data
Train on each data set
Model Evaluation Results ①
Model Evaluation Results ②
Model Evaluation Results ③
Model Evaluation Results ④
Model Evaluation Results ⑤
Model Evaluation Results ⑥
Model Evaluation Results ⑦
Model Evaluation Results ⑧
Model Evaluation Results ⑨
Model Evaluation Results ⑩
Calucuate
Average Accuracy

Figure 1. Overview of the 10-Fold Cross-Validation.

As metrics to evaluate sentences generated by each model, we use BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002), which is often used in machine translation tasks, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004), which is used in summarization tasks, and CR (Content-Rrate), which indicates how many keywords given as input data are included in the generated sentences. If CR is 0.5, it means that the sentence contains the half of the keywords given as input.

3.4 *Information Retrieval-Based Method*

Our IR-based method uses OkapiBM25 (BM25 for short), which is a well-known and commonly used method in the field of information retrieval and is expected to be more accurate than simply using tf-idf values. In this study, we rank the report sentences by the BM25 score based on the input keywords and return the highest ranked report sentences. In other words, this method selects sentences rather than generates them. *Figure 2* illustrates the process of generating sentences based on keywords using BM25. In the IR-based method, the sentence with the highest BM25 score is selected from the training data as output. In a simple way, we need to calculate the BM25 scores of all sentences in the training

[1] http://taku910.github.io/mecab/
[2] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

data. However, the BM25 score is inevitably higher for sentences that contain many words that are input keywords. Therefore, in this study, we first select sentences containing the most input keywords as candidate sentences, calculate the scores of the candidate sentences, and select the sentence with the highest score as the generated sentence.

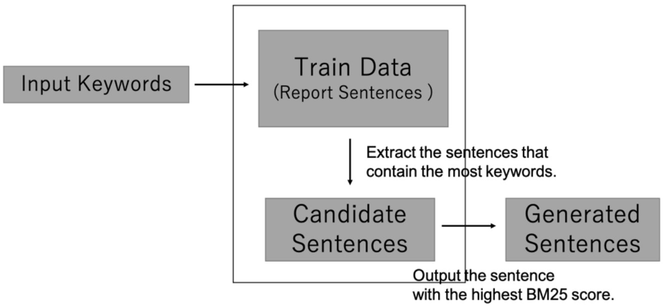

Input Keywords
Train Data
(Report Sentences)
Extract the sentences that contain the most keywords.
Candidate Sentences
Generated Sentences
Output the sentence with the highest BM25 score.

Figure 2. The Process of Sentence Generation by IR-based Method.

3.5 *Seq2Seq-based Method*

Seq2Seq is a model that uses a neural network to convert a sequence data to another sequence data. Seq2Seq is composed of two mechanisms: Encoder and Decoder. LSTM (Long short-term memory) (Gers et al., 1999) is often used for each mechanism. Encoder takes a sequence as input and maps it to a fixed-dimensional vector. The decoder decodes the fixed-dimensional vector output by the encoder to the target sequence. In this experiment, the model is trained to encode the input keywords and decode them into the original report sentences. Here, there are two methods for sentence generation: probabilistic generation (PG) and deterministic generation (DG); PG chooses words according to a probability distribution, and DG uniquely chooses the most probable word. We respectively call the two methods, the Seq2Seq with PG and Seq2Seq with DG methods.

We compare the performance of these methods on generating report sentences. *Figures 3* showss the overviews of the Seq2Seq model during training and evaluation phases, respectively. During training, [KEYWORDS] feeds Instructor's Report Sentence-*i* in the training data and extracts Keywords-*i* that are paired with the Instructor's Report Sentence-*i* using the method described in Section 3.1. Giving the Instructor's Report Sentence-*i* and Keywords-*i* to [SEQ2SEQ] as input, [SEQ2SEQ] adjusts the parameters. By iterating this process, [Trained SEQ2SEQ] is built. When evaluating the model, the Instructor's Report Sentence-*j* is taken from the test data and given to [KEYWORDS]. Keywords-*j* are extracted in the same way as during training. With this Keywords-*j* as input, [Trained SEQ2SEQ] generates the Generated Instructor's Report Sentence-*j*. To evaluate the model, the Instructor's Report Sentence-*j*, Keywords-*j*, and the Generated Instructor's Report Sentence-*j* are given to [JUDGE], and the evaluation score (Score) is calculated according to the metrics: BLEU, ROUGE and CR.

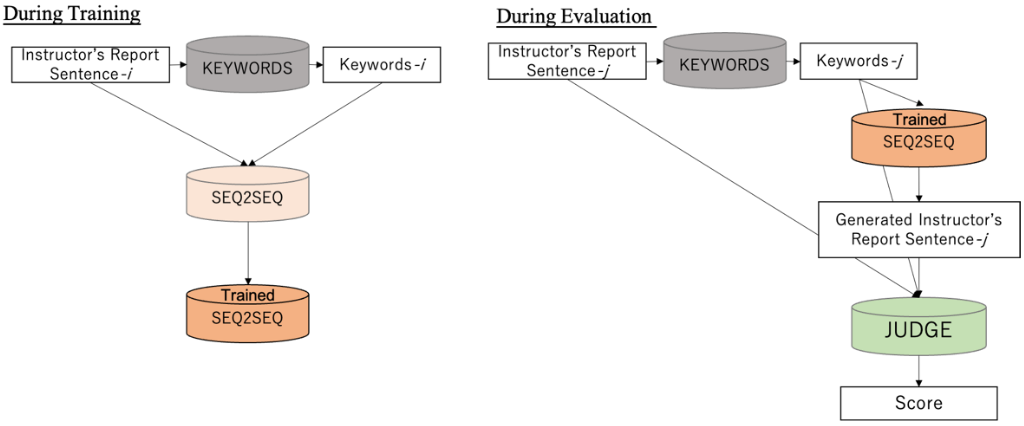


Figure 3. The Seq2Seq Model.

4. Experiments

4.1 *Comparison between IR-based and Seq2Seq-based Methods*

The experimental results are shown in *Figure 4*. For the content-rate, the IR-based method had the highest accuracy, and for the other metrics, the Seq2Seq with DG method had the highest scores. The sentences generated (selected) by the IR-based method tend to include keywords appearing separately, but not connectively, which reduces the rate of n-grams, makes BLEU and ROUGE scores lower, and often generates sentences with unintended contexts. As the BM25 score is higher, the greater number of keywords are included, which selects longer sentences more likely. At the same time, this makes it difficult to generate short and concise sentences. As a result, the Seq2Seq-based method performs better than the IR-based method when generating concise and/or abstract sentences, which are the most part of the test data. On the other hand, the Seq2Seq-based method is not good at generating concrete and complex long sentences, which are well-suited for the IR-based method.

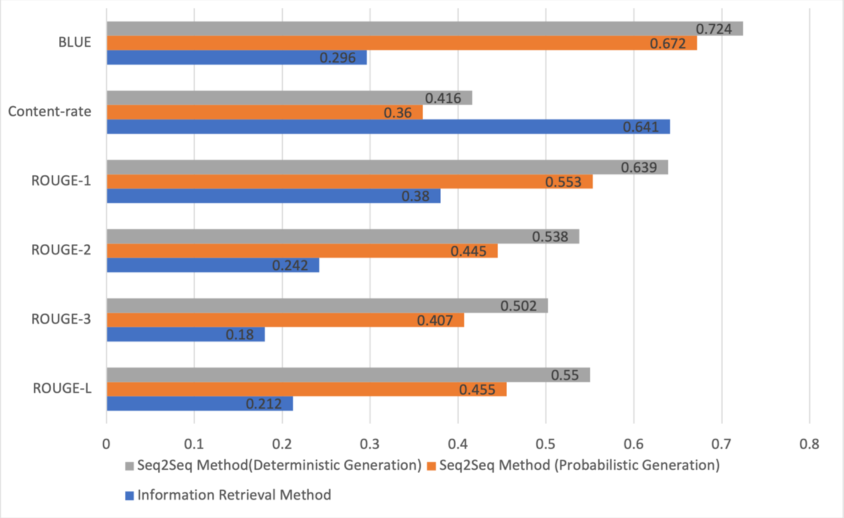


Figure 4. The Comparison Results of the Two Methods.

4.2 *Improving Performance of Seq2Seq-based Method using Attention Mechanism*

As mentioned in the previous section, the Seq2Seq with DG method had the highest accuracy. In this section, we evaluate the effect of Attention mechanism (Vaswani et al., 2017), which is a mechanism that can directly refer to the information of the input sequence at the time of decoding, and can consider the length of the input sequence at the encoder side. Here, this Attention refers to soft Attention. We add the attention mechanism to the Seq2Seq with DG and call the Seq2Seq with Attention method. In contrast, we call the Seq2Seq with DG, but without Attention mechanism, the Seq2Seq without Attention method. We compare the Seq2Seq with and without Attention methods and show the results in *Table 1*. As we can see, the Seq2Seq with Attention method has better performance on ROUGEs, but slightly worse on BLEU. This shows that the Attention mechanisms work effectively in generating report sentences from keywords. Also, *Table 2* shows the specific results of the sentences generated by the Seq2Seq with Attention.

Table 1. *Comparison of Seq2Seq with and without Attention Methods*

	Seq2Seq without Attention	Seq2Seq with Attention
BLEU	***0.724***	0.716
ROUGE-1	0639	***0.725***
ROUGE-2	0.538	***0.620***
ROUGE-3	0.502	***0.565***
ROUGE-L	0.550	***0.608***

Table 2. *Report sentences Generated by the Seq2Seq with Attention*

Keywords	Generated sentence
単語　ミス　有り (word errors exist)	単語のミスが少し有りますが、少しずつ理解してくれています （There are a few word errors, but you’re slowly getting the hang of it.）
授業　集中　できる （class concentrate can）	授業中、集中して問題に取り組んでくれました （You concentrated on the problem during the class）

5. Conclusion

This paper discussed report generation models for instructors who have to write reports on learning status of their students. Using the models, the instructors can obtain student progress report by simply inputting keywords instead of writing many sentences. This can greatly reduce instructors’ burden on generating report sentences. In this study we proposed two sentence generation methods: the IR-based method using OkapiBM25 and the Seq2Seq-based method, and compared their performance by conducting experiments on a dataset consisting of about 200,000 report sentences written by the instructors at cram schools. Experimental results show the Seq2Seq-based method outperforms the IR-based method and adding the attention mechanism to the Seq2Seq-based method had effects on improving the performance of generating sentences. However, our keyword-based sentence generation method using the Seq2Seq-based method still has some weak points, especially on generating concrete and detail sentences, which will greatly improve the usability of this model. We will commit to tackle this problem and report it elsewhere in near future. In addition, the system that actually uses Seq2Seq with Attention is currently being highly evaluated and used by people involved in the cram schools.

Acknowledgements

This work was supported in part by e-sia Corporation and by Grant-in-Aid for Scientific Research proposal numbers (JP21H00907, JP20H01728, JP20H04300, JP19KK0257, JP18K18656). We would like to express our deepest gratitude to them.

Reference

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., & Gatford, M. (1995). Okapi at TREC-3. Nist Special Publication Sp, 109, 109.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. arXiv preprint arXiv:1409.3215.

Gutl, C., Lankmayr, K., Weinhofer, J., & Hofler, M. (2011). Enhanced Automatic Question Creator--EAQC: Concept, Development and Evaluation of an Automatic Test Item Creation Tool to Foster Modern e-Education. *Electronic Journal of e-Learning*, *9*(1), 23-38.

Liang, G., On, B. W., Jeong, D., Kim, H. C., & Choi, G. S. (2018). Automated essay scoring: A siamese bidirectional LSTM neural network architecture. *Symmetry*, *10*(12), 682.

Lu, C., & Cutumisu, M. (2021). Integrating Deep Learning into An Automated Feedback Generation System for Automated Essay Scoring. EDM 2021

Malik, A., Wu, M., Vasavada, V., Song, J., Coots, M., Mitchell, J., ... & Piech, C. (2021). Generative Grading: Near Human-level Accuracy for Automated Feedback on Richly Structured Problems. EDM 2021

Parab, A. K. (2020). Artificial Intelligence in Education: Teacher and Teacher Assistant Improve Learning Process. International Journal for Research in Applied Science & Engineering Technology, 8(11), 608-612.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation. ACL 2002, 311-318.

Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. Text summarization branches out, 74-81.

Gers, F. A., Schmidhuber, J., & Cummins, F. (1999). Learning to forget: Continual prediction with LSTM. Neural computation, 12(10), 2451-2471.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.

Gaze- and Semantics-aware Learning Material to Capture Learners' Comprehension Processes

Akio OKUTSU[a*], Yuki HAYASHI[b] & Kazuhisa SETA[b]
[a]*College of Sustainable System Sciences, Osaka Prefecture University, Japan*
[b]*Graduate School of Humanities and Sustainable System Sciences, Osaka Prefecture University, Japan*
*okutsu@ksm.kis.osakafu-u.ac.jp

Abstract: Learners' comprehension processes and their states of knowledge can be used to analyze and facilitate learning. The purpose of this study is to develop a system that records the learning processes to estimate the learners' comprehension processes with respect to the target domain-knowledge and state of knowledge from their gaze on the screen when viewing the learning material. Gaze is often interpreted as a part of one's thinking processes. In this study, we developed learning material that records learners' learning processes by measuring their gaze to determine the semantics they tried to understand. In addition, we created an authoring system for developing the learning material.

Keywords: Gaze behaviors, semantic network, learning material, comprehension processes

1. Introduction

It is essential for learners to recognize their own state of knowledge and to self-regulate their processes of understanding in order to learn as effectively as possible (Mayer, 2014). For example, when learning from a textbook, learners need to establish semantic connections with prior knowledge (Chi, & Wylie, 2014). However, these kinds of metacognitive techniques can be difficult for immature learners to acquire (Kayashima, Inaba, & Mizoguchi, 2008). To encourage the use of these techniques, a learning support system must identify the learners' comprehension processes and state of knowledge in detail. Then the system can provide the learners with adaptive feedback and may also reveal the relationship between learning style and academic achievement. Using this analysis, the system may be able to estimate the comprehension level and provide information such as posing questions to encourage understanding and drawing the learners' attention to information they need to pay further attention to. The significance of such learning support is that it is suitable to each individual's learning process. In contrast, conventional learning support systems encourage learners to engage in different problem-solving situations to diagnose their state of understanding.

Many studies have focused on gaze as a way to understand the implicit nature of a learners' thinking processes. Gaze measurement has been used effectively in various cognitive processing analyses (Ohno, 2002) because gaze behaviors reflect a part of the thought processes without interfering with higher-order human cognitive processing. Furthermore, gaze measurement can be combined with other analytical methods and has been recognized as a promising method to approach cognitive and metacognitive learning processes that cannot be captured by learners' self-reported cognitive judgments (Roderer & Roebers, 2014; Hayashi, Seta, & Ikeda, 2018).

Several studies have used gaze behaviors in learning contexts to reveal the characteristics of learners. Antonietti, Colombo, and Nuzzo (2015) found that the frequency of a learner's attention indicates what they consider important in the learning material, while the length of their attention indicates their interest in the information, and exceedingly long attention indicates difficulties in cognitive processing. Mason, Pluchino, and Tornatora (2013) reported that the sequence in which information is presented in text and images reveal the cognitive processes of integrating with prior knowledge. Furthermore, the order in which learners pay attention to a text and the corresponding image may indicate the learners' cognitive processes of constructing organized knowledge and mental models.

A high frequency of close attention by gazing back and forth between the text and image may be related to academic achievement (Mason, Pluchino, & Tornatora, 2015).

Many studies have attempted to capture learners' thought processes or higher-order cognitive processing using gaze information on learning materials; however, these studies only analyzed characteristics of gaze behavior and comprehensive states and processes from visual components (e.g., text, figures, tables) of the information. The studies lack a framework to clarify the characteristics based on the specific knowledge content (semantic structure) embedded in the learning materials. The researchers gave their semantic interpretation of learners' gaze behaviors on the learning material manually, but this is not practical. Therefore, in order to accumulate further academic knowledge on gaze analysis, it is necessary to build a framework that can handle the semantics of knowledge content in addition to the visual components.

We previously developed gaze- and semantics-aware learning material to adaptively pose questions on the basis of the learner's attention. The material can be used to examine the correlation between the learner's gaze on the learning material and the semantics of the information (Muroya, Seta, & Hayashi, 2021). To make this method applicable to a variety of learning domains, in this paper, we propose an authoring system for creating gaze- and semantics-aware learning materials. Then, using the developed learning materials, we measure learners' gaze information on science-related learning material and investigate the system's capability to analyze and estimate the learners' learning processes using the semantics of the knowledge content. For this purpose, we use learning material containing contradictory information, which would cause cognitive conflicts during the learning processes if a learner tries to integrate knowledge rather than merely store it. The characteristics of measured gaze information on the contradictory knowledge may be differentiated by the undertaken cognitive processes. Thus, we investigate the system's capability of measuring such characteristic gaze information concerning knowledge content and capturing the degree of schema formulation.

2. Functional Requirements

To support learners constructing well-organized knowledge in textbook-based learning, this study focuses on the learners' gaze on the learning material to estimate their comprehension processes and knowledge state. Our final goal is to build a framework for interpreting gaze behaviors and estimating the learner's knowledge state by representing the semantic relations of the learning materials in the system.

A semantic network is a well-known knowledge representation method that defines semantic relations between concepts using nodes (concepts) and links (relations). The network enables semantic connections between concepts in the learning material to be represented in a machine-readable manner, including information that is not explicitly mentioned on the textbook but should be learned in order to construct well-organized knowledge.

We establish a set of knowledge in the semantic network for a particular area of the learning material as the area of interests (AOIs) in advance. This results in gaze- and semantics- aware learning material as the system can detect the semantics of the information the learner is gazing at based on the eye-tracking device. When used with such learning materials, the system can record learning processes, such as who pays how much attention to which information or the order in which the learner looks at the information, and stores this data as a learning log. By analyzing the learning log, it may be possible to trace the comprehension process and estimate the state of knowledge of the learner.

The following three functional requirements must be satisfied for this system to be feasible:

- **Requirement 1.** *An authoring support function for creating gaze- and semantics-aware learning materials*: In order for the system to trace the comprehension processes from the learners' gaze behaviors, the system must have the information of what knowledge is represented at what area on the learning material. Therefore, an authoring function is needed which enables the author to create semantic networks for the target learning material and set AOIs on the material corresponding to the semantic structure of knowledge, thereby creating gaze- and semantics- aware learning materials.
- **Requirement 2.** *Functions for measuring and recording gaze information during learning*: Semantics-aware gaze measurement and recording functions need to be used in conjunction with the

learning materials stated above. The functions would capture gaze information such as the learner's gaze time and the order in which the learner focused on the content, in addition to the media properties they focused on (text, figures, tables, etc.) and record them as a learning log.

- **Requirement 3.** *Learning support functions based on estimated learner's knowledge state by learning records*: Support functions are needed based on estimating the learner's comprehension process and the state of knowledge from the learning log. Then the system needs to provide adaptive feedback, such as presenting information that may be not fully understood or guiding the eyes on the learning material to promote structural understanding (e.g., highlighting the learning contents).

In this paper, we aim to fulfill the Requirements 1 and 2, which are needed to realize Requirement 3 in the future.

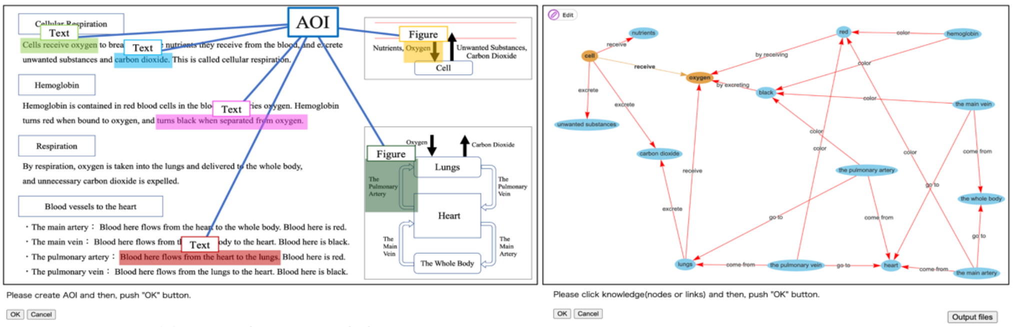
3. Proposed Systems

In this study, we developed two systems, an authoring system for gaze- and semantics-aware learning materials and a semantics-aware gaze measurement system to satisfy Requirements 1 and 2 described in the previous section.

3.1 Authoring System for Gaze- and Semantics-Aware Learning Materials

Figure 1 shows the interface of the developed authoring system. The system is implemented as a web application using JavaScript and HTML and runs in widely used browsers such as Chrome, Safari, and Edge. The interface consists of a learning material area (Fig. 1(a)) and a semantic network editing area (Fig. 1(b)).

The author uses the system by inputting the image file of the learning material (e.g., png format) and the file of the corresponding semantic network (xml format) that represents the semantic structure of knowledge. In the semantic network area, the author can freely edit and extend the displayed semantic network (e.g., create/delete nodes and links) accordingly for the learning material. In the learning material area, the author can drag the mouse to create a rectangular area indicating the AOI that captures the target of the learner's gaze from the eye-tracking device. Then the author selects the corresponding semantic structure of the knowledge to the AOI by clicking nodes and links in the semantic network editing area. After the selection, the system requires the author to select the media property of the AOI from 'Sentence,' 'Figure,' or 'Table.' By repeating these operations, a semantic network is linked to several areas in the learning material. Here, the system allows the author to associate different media in the learning material with the same knowledge in the semantic network. For example, if the author creates AOI on the text about "Cells receive oxygen", the corresponding semantic structure of the knowledge ("cell" node, "oxygen" node, and "receive" link) should be chosen in the semantic network, and 'Sentence' should be selected as the media property. On the other hand, if the author creates AOI on the area in a figure about the same information, the same semantic structure of the knowledge ("cell" node, "oxygen" node, and "receive" link) should be chosen in the semantic network, while 'Figure'



Cells receive oxygen to brea[...] nutrients they receive from the blood, and excrete unwanted substances and carbon dioxide. This is called cellular respiration.

Hemoglobin is contained in red blood cells in the blo[...]ies oxygen. Hemoglobin turns red when bound to oxygen, and turns black when separated from oxygen.

By respiration, oxygen is taken into the lungs and delivered to the whole body, and unnecessary carbon dioxide is expelled.

・The main artery： Blood here flows from the heart to the whole body. Blood here is red.
・The main vein： Blood here flows from t[...]ody to the heart. Blood here is black.
・The pulmonary artery： Blood here flows from the heart to the lungs. Blood here is red.
・The pulmonary vein： Blood here flows from the lungs to the heart. Blood here is black.

Please create AOI and then, push "OK" button.

Please click knowledge(nodes or links) and then, push "OK" button.

(a) Learning Material (b) Semantic Network

Figure 1. Authoring System for Gaze- and Semantics-aware Learning Materials.

should be selected as the media property.

Finally, the system outputs two files, the final state of edited semantic networks and AOI information that includes the ID of each AOI, position in the learning material, size, linked knowledge, and media property.

3.2 Semantics-Aware Gaze Measurement System

Figure 2 shows the developed gaze measurement system. The system is implemented as a C# form application and works with a screen-based eye tracker (Tobii Eye Tracker 4C with analytical license) to measure the learner's gaze behaviors. The system manages the learning material layer displayed to the learner (Fig. 2(a)) and the semantic network layer (Fig. 2(b)) that connects each AOI to the corresponding semantic structure of the knowledge.

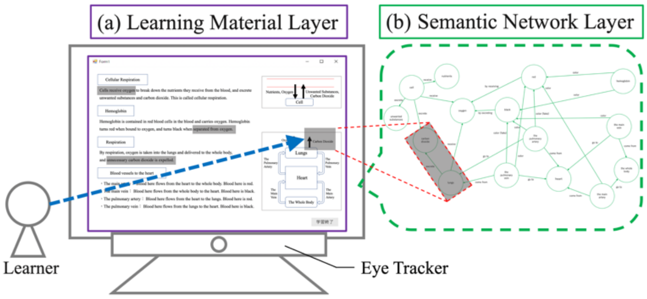


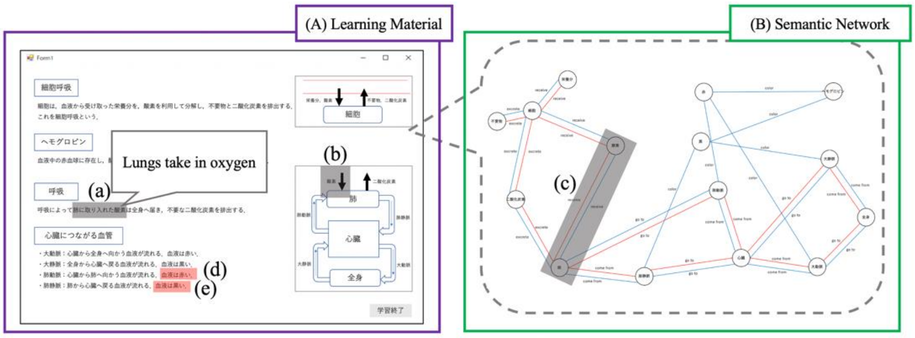
Figure 2. Semantics-aware Gaze Measurement System.

Before learning, the system requires the following three files to be input: the image file of the learning material (e.g., png format), the semantic network file (xml format), and the AOI information file (xml format) output by the authoring system described in Section 3.1. Then, the image of the input learning material appears on the interface. Although AOIs are set on the learning material on the basis of the input file (i.e., shaded area in Fig. 2(a)), the AOIs are invisible on the interface so as not to interfere with learning.

Learners carry out their learning using the learning material. During the learning processes, the system monitors the learner's gaze behavior, whether the gaze falls within certain AOI regions in each frame, and detects the corresponding knowledge in the semantic network. The information measured by the system is saved as a learning log in which each line includes the ID of the gaze target AOI, gaze state (start/stop), timestamp (ms), corresponding knowledge, and media property.

4. Experimental Study

To evaluate the feasibility of the proposed framework, we conducted the experimental study whether the developed systems fulfill the functional requirements 1 and 2 and confirm its potential. We created a learning material about blood circulation taught in junior high school science classes. The learning material included text and figures described in Japanese (Fig. 3(A)). We prepared 15 concepts and 34


細胞呼吸

細胞は、血液から受け取った栄養分を、酸素を利用して分解し、不要物と二酸化炭素を排出する。これを細胞呼吸という。

ヘモグロビン

血液中の赤血球に存在し、

呼吸

呼吸によって肺に取り入れた酸素は全身へ届き、不要な二酸化炭素を排出する。

心臓につながる血管

- 大動脈：心臓から全身へ向かう血液が流れる。血液は赤い。
- 大静脈：全身から心臓へ戻る血液が流れる。血液は黒い。
- 肺動脈：心臓から肺へ向かう血液が流れる。血液は赤い。
- 肺静脈：肺から心臓へ戻る血液が流れる。血液は黒い。

Figure 3. Learning Material and Its Semantic Network used for Learning.

links as the semantic network and set 29 AOIs using the authoring system. In the semantic network (Fig.3(B)), the blue and red links indicate the media property of the corresponding set of AOIs (blue: text, red: figure). For example, if the learner is gazing at the text “lungs take in oxygen (Fig. 3(A-a))” in the learning material, the system can detect that the learner is paying attention to the corresponding knowledge (Fig. 3(B-c)). The system distinguishes each AOI from the media property even if the related knowledge is the same. Thus, when the learner is gazing at the area in a figure about the information “lungs take in oxygen” (Fig. 3(A-b)), the system can grasp the corresponding knowledge (Fig. 3(B-c)) and record it separately from the text (Fig. 3(A-a)).

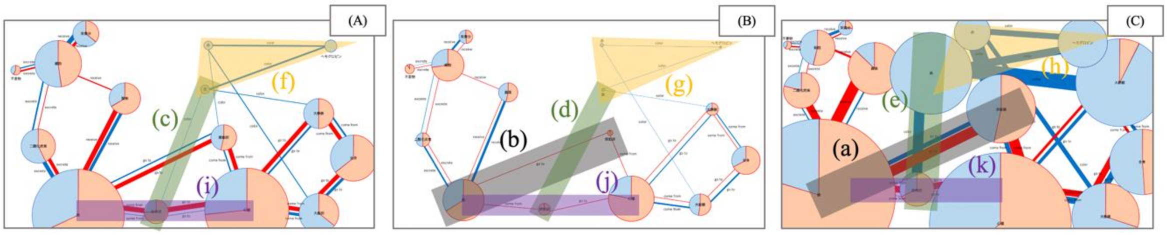


Figure 4. Visualization of Learning by each Learner.

In this experiment, the learning material includes incorrect information about the color of blood flowing through pulmonary arteries and veins as the text contents in the learning material (Fig. 3(A-d) and 3(A-e)). The objective of including incorrect information is to investigate whether the difference appears in gazing behaviors when the learner notices and causes cognitive conflicts from the inconsistent information.

Three junior high school students (learners A, B, and C) participated in learning using the semantics-aware gaze measurement system. We did not tell the learners that the material included incorrect information as mentioned previously.

Figures 4(A), 4(B), and 4(C) represent the visualized semantic network based on learning logs that recorded the learning processes of the three learners, respectively. The size of the circle (concept) and the thickness of the line (link) represent the total amount of time spent paying attention to the corresponding AOIs on the learning material. Here, the blue nodes and links indicate the rate of knowledge focused on textual information and the red ones represent the same for graphic information (i.e., figures on the learning material).

The results indicate that learner C spent the longest time among the three learners and paid attention to a wider variety of information. In addition, we observed the different tendencies of learners’ focused media property. For example, while learner C paid attention to the information “the pulmonary artery is connected to the lung” in both the textual and graphic information (Fig. 4(C-a)), learner B only paid attention to the graphic information and the attention time was quite shorter than that of learner C (Fig. 4(B-b)).

In the post-learning interview, only learner C commented that she noticed the inconsistent knowledge (shown in Fig. 3(A-d) and 3(A-e)). The nodes and links corresponding to inconsistent information on the semantic network are “the color of the blood flowing through the pulmonary veins (Fig. 4(A-c), 4(B-d), 4(C-e)),” “the color change of hemoglobin (Fig. 4(A-f), 4(B-g), 4(C-h)),” and “the way the pulmonary veins connecting to the lungs and heart (Fig. 4(A-i), 4(B-j), 4(C-k)).” From the three visualization diagrams, we found that learner C paid more attention to the inconsistent information than the other two learners. In addition, we found three characteristic gaze events from the timestamp information in learner C’s log: (i) the ratio of total gazing time to the inconsistent information was larger than that of the other two learners, (ii) after she paid attention to the incorrect information, she gazed at the related information and (iii) she paid attention to both textual and graphic information represented the same knowledge. The ratio of gazing time (i) suggests that the learner may have been facing a cognitive conflict when she noticed the inconsistent information. She tried to structurally understand the learning contents by comparing and integrating them with her knowledge structure. To resolve the cognitive conflict, the learner would attempt to check the semantic connection between the inconsistent information one at a time, as suggested by (ii). (iii) suggests that the learner could recognize that the text and figures contained the same information, and she processed information systematically by contrasting the text and figures.

5. Concluding Remarks

In this study, we confirmed the proposed framework captures characteristics of each learner at the semantic level to some extent. It has potential towards learning contents oriented active learning analytics and supports. That is, in previous studies, the researchers manually analyzed the semantics of the knowledge in the learning material against information derived from gaze, such as cognitive and metacognitive processes of learners (e.g., Mudrick, Azevedo, & Taub, 2019) and the learning styles. In contrast to the conventional analysis, we proposed an authoring system for creating the gaze- and semantics-aware learning materials (Requirement 1) and a measurement system (Requirement 2) that can automatically capture several semantic relationships of information represented in the learning material (e.g., contents containing the same information, other semantically connected information, and inconsistent information).

Learning is a complex and implicit activity requiring both cognitive and metacognitive processing. In order for the information system to adaptively intervene in learners' knowledge construction, the system must be able to capture their comprehension processes. The framework proposed in this paper is a promising approach toward this purpose.

In this study, we investigated the feasibility of the proposed systems through an initial operation verification by three learners. In the future, we plan to implement the learning materials in existing learning support systems, such as the one developed by Aburatani et al. (2019), to investigate learners' comprehension processes by analyzing learning logs. In addition, we need to consider the functions of an intelligent learning support system (Requirement 3) that provides feedback based on the learners' estimated comprehension processes and the state of knowledge.

References

Aburatani, T., Seta, K., Hayashi, Y., & Ikeda, M. (2019). Implementation of a System for Diagnosing Learning Behaviors of Thinking between the Lines Based on Presentation Design Tasks. *IEICE Transactions on Information and Systems*, Vol.J102-D, No.4, 359-363 (in Japanese).

Antonietti, A., Colombo, B., & Nuzzo, C. D. (2015). Metacognition in self-regulated multimedia learning: Integrating behavioural, psychophysiological, and introspective measures. *Learning, Media, and Technology*, 40, 187-209.

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*, 219-243.

Hayashi, Y., Seta, K., & Ikeda, M. (2018). Development of a support system for measurement and analysis of thinking processes based on a metacognitive interpretation framework: a case study of dissolution of belief conflict thinking processes, *Research and Practice in Technology Enhanced Learning*, 13-21.

Kayashima, M., Inaba, A., & Mizoguchi, R. (2008). A Framework of Difficulty in Metacognitive Activity. *Transactions of Japanese Society for Information and Systems in Education*, 25(1), 19-31.

Mason, L., Pluchino, P., & Tornatora, M. C. (2013). Effects of picture labeling on science text processing and learning: Evidence from eye movements. *Reading Research Quarterly*, 48, 356-389.

Mason, L., Pluchino, P., & Tornatora, M. C. (2015). Eye-movement modeling of integrative reading of an illustrated text: Effects on processing and learning. *Contemporary Educational Psychology*, 41, 172-187.

Mayer, R. E. (2014). Cognitive theory of multimedia learning. *The Cambridge handbook of multimedia learning, Cambridge University Press*, 43-71.

Mudrick, N. V., Azevedo, R., & Taub, M. (2019). Integrating metacognitive judgments and eye movements using sequential pattern mining to understand processes underlying successful multimedia learning. *Computers in Human Behavior*, Vol.96, 223-234.

Muroya, D., Seta, K., & Hayashi, Y. (2021). Semantically Enhanced Historical Cartoons Promoting Historical Interpretation. *Information and Technology in Education and Learning*, 1(1), Reg-p002.

Ohno, T. (2002). What Can Be Learned From Eye Movement?: Understanding Higher Cognitive Processes From Eye Movement Analysis. *Cognitive studies: bulletin of the Japanese Cognitive Science Society*, Vol.9, No.4, 565-579 (in Japanese).

Roderer, T., & Roebers, C. M. (2014). Can you see me thinking (about my answers)? Using eye-tracking to illuminate developmental differences in monitoring and control skills and their relation to performance. *Metacognition and Learning*, 9, 1-23.

Learning the Condition of Addition and Subtraction Word Problems by Problem-Posing Based on Representation Conversion Model

Yusuke HAYASHI[a], Natsumi TSUDAKA[a], Kengo IWAI[a] & Tsukasa HIRASHIMA[a]
[a]*Hiroshima University, Hiroshima, JAPAN*
*hayashi@lel.hiroshima-u.ac.jp

Abstract: Understanding arithmetic word problems can be said as a structural understanding of the relationship between linguistic and mathematical representations. The goal of this study is to make learners understand the conditions of addition or subtraction word problem. This study proposes an exercise of the conversion among linguistic, mathematical, and graphic representations. Monsakun-TapeBlock is a learning environment to conduct the exercise. The effectiveness of these exercises is suggested through the experimental use in a public elementary school. The learners came to explain the relation between the linguistic and the mathematical representation with the quantity roles and judge valid arithmetic word problems better.

Keywords: Arithmetic Word Problem, meaningful approach, Problem-posing, Graphic representation

1. Introduction

Solving an arithmetic word problem can be said to read sentences, extract the quantity relationship in them, represent it as a mathematical representation, and derive unknown numbers (Mayer, 1992). Many researchers have already investigated solving process of the word problems, and they have divided the process into two sub-processes: (1) comprehension phase and (2) solution phase (Cummins et al., 1988; Hegarty et al., 1995; Heffernan and Koedinger, 1998; Pólya, 1945; Riley et al., 1983). They have also pointed that the comprehension phase plays a major role in the difficulty of the word problems. In this phase, a learner is required to interpret the representation written by words and to create quantitative relationships. Here, several researchers assumed that the product of the comprehension phase is a representation that is connecting "problem (conceptual) representation" and "quantitative representation " (Koedinger & Nathan, 2004; Nathan et al., 1992; Reusser, 1996).

There are two general approaches in the comprehension phase: a short-cut approach and a meaningful approach (Hegarty et al., 1995). In the short-cut approach, the problem solver attempts to select the numbers in the problem and critical relational terms (such as "more" and "less" which imply an operation) and make a numeric formula. On the other hand, in the meaningful approach, the problem solver translates the problem statement into a mental model to the situation described in the problem statement. The mental model becomes the basis for the solution phase.

The meaningful approach has two learning methods. One is problem representational technique in problem-solving, and the other is problem-posing. Problem representational technique provides learners with schematic diagrams. Schematic diagrams represent problem schema depending on problem types, e.g., change, combine and compare in addition and subtraction word problems (Riley et al., 1983). Schema-based strategy is better than the traditional approaches. Several investigations have confirmed that learning by problem-posing in conventional classrooms is a promising activity in learning mathematics (Silver & Cai, 1996; English, 1998). Since learners are usually allowed to pose several kinds of problems, and they can make an extensive range of problems, it is difficult for teachers to complete assessment and feedback for the posed problems in classrooms practically.

This study proposes the addition of problem representational technique in problem-posing with the learning environment for learning arithmetic word problems by sentence-integration "Monsakun" (Hirashima et al., 2007; Hirashima et al., 2008; Hirashima & Kurayama, 2011) as a reflection after problem-posing. In Monsakun, instead of writing sentences freely, learners pose arithmetic word problems as the combination of the provided sentences. The task learners perform in Monsakun is to pose arithmetic word problems satisfying the required condition; numeric formula and the type of story. This task requires learners to make a mental model of a story that can be solved by the provided numeric formula by themselves. This study uses the graphic representation to support this process in which learners can check the validity of the posed problems.

This paper has the following structure. Section 2 proposes a graphic representation of numerical relationships in the one-step arithmetic word problem and shows the learning environment in which learners convert a problem statement to a numeric formula through the graphic representation. Section 3 reports a case study of the use of the learning environment in an elementary school. Section 4 concludes this paper.

2. Representation Conversion Model for Arithmetic Word Problems

This study models arithmetic word problems as the relationship between the linguistic, graphical, and mathematical representations shown in Fig. 1, based on the Triplet-structure model (Hirashima et al., 2014). The linguistic representation of an arithmetic word problem is the problem statement, for example, “Three flowers were in bloom. Two flowers bloomed. Five flowers are in bloom.” In the model, the linguistic representation is the composition of three simple sentences expressing a quantity relationship to clarify the roles of the quantities in the statement. Each simple sentence has a quantity (in this figure, the number of flowers), and each quantity has a role depending on the meaning of the statement. In the case of Figure 1, the quantity of three flowers has the role of a *start* quantity, the quantity of two flowers has the role of a *change* quantity, and the quantity of five flowers has the role of *end* quantity in the story.

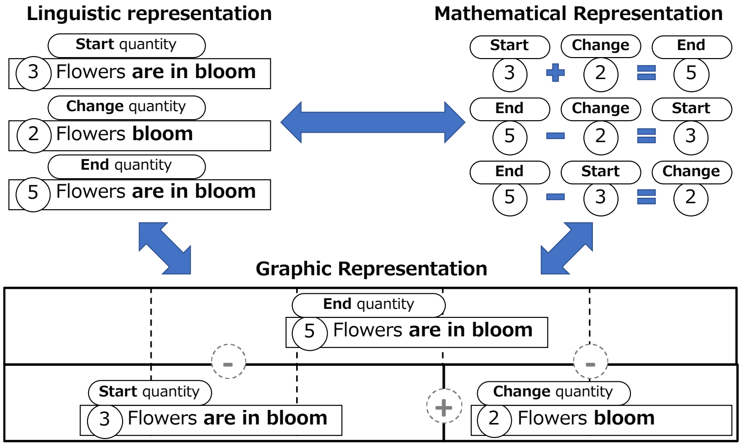


Figure 1. The Relationship between the Linguistic, Graphical, and Mathematical Representations.

This study proposes “Tape-block” as the graphic representation of the relation among the quantities. The bottom of Fig. 1 shows the proposed graphic representation based on part-part-whole schema and drawing (Resnick, 1983; Willis & Fuson, 1988). This representation itself describes the relation among three quantities, in which the top quantity is the sum of the two bottom quantities. This also describes the difference between the top quantity and a bottom quantity in the other bottom quantity. This means that a numerical relation in this type can be converted to three types of equality: one addition and two subtractions shown at the top-left in Fig. 1. The correspondence between the graphical and the mathematical representation is by the magnitude of quantities. The largest number in the mathematical representation must be located at the top of the graphic representation. On the other hand, the correspondence between the graphical and the linguistic representation is by the meaning of the quantities. The resultant quantity in the linguistic representation must be located at the top of the graphic representation. These correspondences can explain the correspondence between the linguistic and the mathematical representation.

Figure 2 shows the classification of arithmetic word problems and the comparison between the schematic drawing and Tape-block diagram. There are four types of arithmetic word problems: Put-together (combine), change-get-more, change-get-less, and compare. Schematic drawings define four different drawings depending on the type of arithmetic word problems. Tape-block diagram uses the same diagram for all the type of arithmetic word problems. This diagram also implies the algebraic operators between the quantities. For example, in Change-get-more, the addition operator between *Start* and *Change* means the equation: *Start* quantity + *Change* quantity = *End* quantity. Based on the diagram, learners can consider valid equations among three quantities without algebraic manipulations, such as transposition of terms, which elementary school students have not learned.

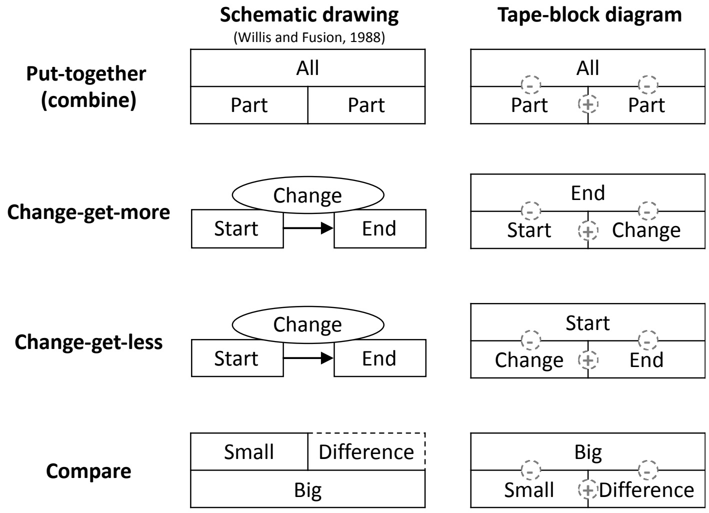


Figure 2. The Classification of Arithmetic Word Problems.

Based on the model of the conversion among linguistic, mathematical, and graphic representation with a Tape-block diagram, we developed a learning environment named Monsakun-TapeBlock. This has the exercises of all the types of conversion of linguistic, mathematical, and graphic representation and quantity role assignment in each representation. Here, due to the limitation of the space, we explain the problem-posing process in Monsakun-TapeBlock as an example. Fig.3 to Fig.6 shows the screenshot of the conversion from mathematical representation to linguistic one through graphic one and quantity role assignment in the process.

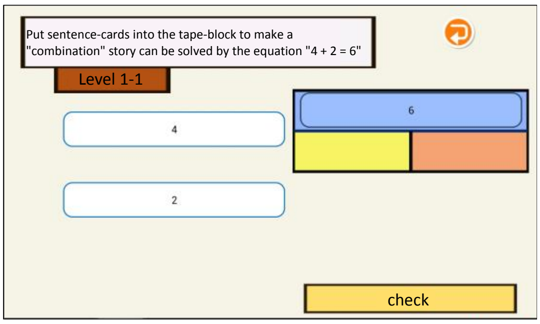
3. Experimental Use in an Elementary School

We conducted the experimental use of this learning environment and measured the effectiveness of it.

The research questions in this experimental use are the followings:

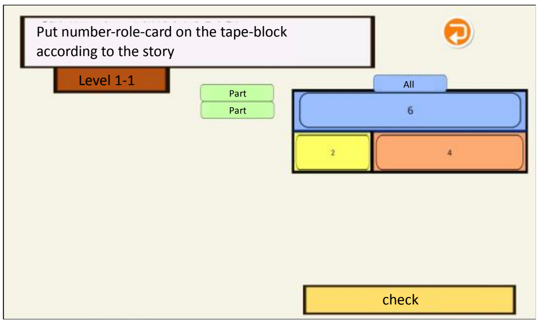
RQ1: Do the students improve the understanding of addition and subtraction word problems?

RQ2: Do the students accept the exercise on Monsakun-TapeBlock?



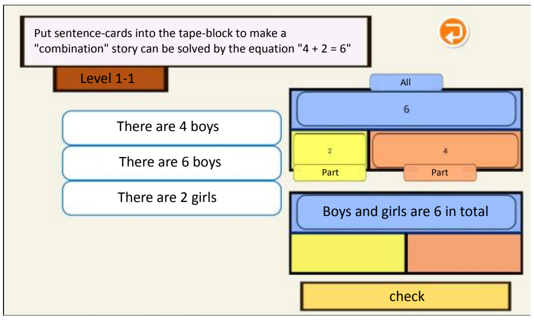
Put sentence-cards into the tape-block to make a "combination" story can be solved by the equation "4 + 2 = 6"

Figure 2. Conversion from Mat Hematical Representation to Graphic One.



Put number-role-card on the tape-block according to the story

Figure 3. Quantity Role Assignment to Numbers in Graphic Representation.



Put sentence-cards into the tape-block to make a "combination" story can be solved by the equation "4 + 2 = 6"

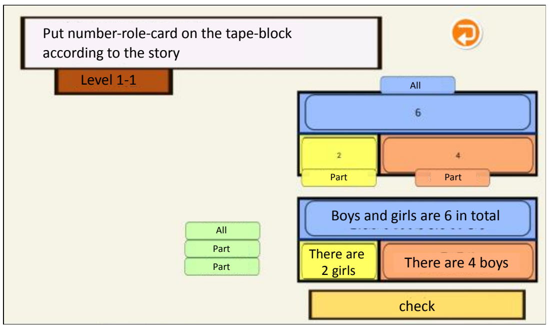
There are 4 boys

There are 6 boys

There are 2 girls

Boys and girls are 6 in total

Figure 4. Conversion from Graphic Representation to a Linguistic One.



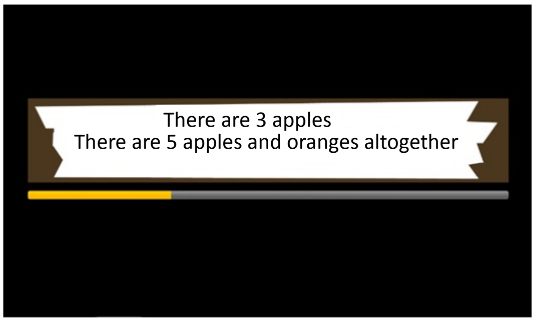
Put number-role-card on the tape-block according to the story

Boys and girls are 6 in total

There are 2 girls

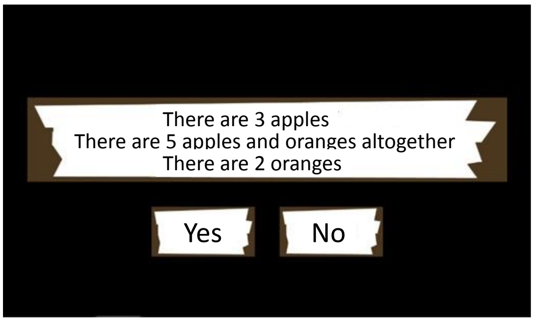
There are 4 boys

Figure 5. Quantity Role Assignment in Graphic Representation with Quantity Propositions.



There are 3 apples
There are 5 apples and oranges altogether

Figure 6. A Screenshot of the First Step in the Priming Test.



There are 3 apples
There are 5 apples and oranges altogether
There are 2 oranges

Figure 7. A Screenshot of the Last Step in the Priming Test.

To answer these research questions, we conducted pre/post-tests and questionnaires in addition to the exercise on Monsakun-TapeBlock. Eighty-eight students in third-grade public elementary school students use Monsakun-TapeBlock in two units of lessons. The procedure is the following:

1. a quick review of addition and subtraction word problems (5 min).
2. pre-test (15 min).
3. exercise on Monsakun-TapeBlock (50 min).
   firstly problem-posing, secondary the conversion from linguistic to graphical, and lastly the conversion from graphical to mathematical representations to check the validity of the posed problems.
4. post-test (15 min).
5. questionnaire (5 min).

To answer RQ1, we develop the structure comprehension test for the pre/post-tests named “priming test.” The priming test asks learners whether the shown story is valid or not. For example, “Jon has 3 apples. Jon and Bill have 8 apples altogether. Bill has 5 apples.” is valid.

On the other hand, “Jon has 3 apples. Jon and Bill have 5 apples altogether. Bill has 8 apples.” and “Jon has 3 apples. Jon and Bill have 8 apples altogether. Bill has 5 oranges.” are not valid. An item shows firstly two sentences and then shows the last sentence. This task requires the understanding of the conditions of addition or subtraction word problem. During the first part, if learners can predict the last sentence, the learners can quickly answer the validity of the story. If they start to think after the display of the last sentence, they take much time to answer. **Error! Reference source not found.** and 8 show the screenshots of it. The progress bar shows the time to display the last sentence. When the bar reaches the right end, the last sentence is displayed shown in **Error! Reference source not found.**. This test has 13 items, including four kinds of story: put-together, change-get-more, change-get-less, and compare.

We answer partially yes to RQ1: “Do the students improve the understanding of addition and subtraction word problems?” is yes. Table 1 shows the result of the pre/post-test. We analyze the difference of score and time between pre and post-test with the Wilcoxon signed-rank test. There is not a significant difference in the score. On the other hand, there is a significant difference in time. This means that they keep improving the understanding of addition and subtraction word problems in terms of the speed of thinking. Unfortunately, they did not get worse and improve in the correctness.

Table 1. *The Result of Pre/Post-Test*

	Pre mean (sd)	Post mean (sd)	p-value
Score (full score is 13)	10.33 (2.32)	10.31 (2.16)	0.9468
Average time per item (sec)	5.77 (3.38)	4.62 (2.65)	0.0007

To answer RQ2, we carried out the questionnaire includes the following:

1. Do you like to study mathematics?
2. Did you use Monsakun-TapeBlock easily?
3. Did you enjoy using Monsakun-TapeBlock?
4. Do you think posing problems in Monsakun-TapeBlock is good for studying mathematics?
5. Do you think quantity role assignment in Monsakun-TapeBlock is good for studying mathematics?
6. Is it easy for you to assign quantity roles in Monsakun-TapeBlock?

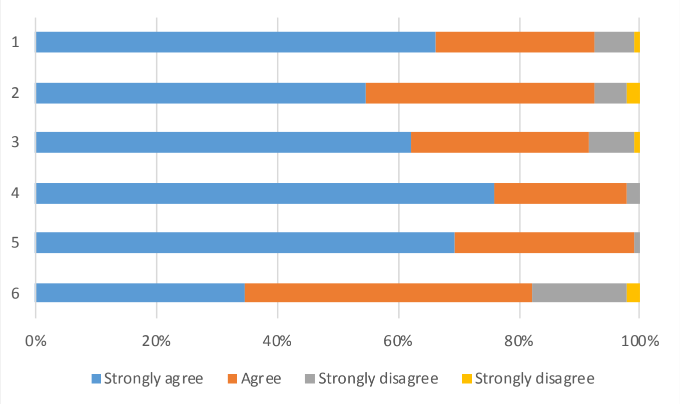


Figure 8. The Result of the Questionnaire

We answer yes to the RQ2: “Do the students accept the exercise on Monsakun-TapeBlock?” ***Figure 8*** shows the result of the questionnaire. From the question 1-3 most of the student like to study mathematics and enjoy the exercise on Monsakun-TapeBlock with easy use. From the question 4-5, most of them enjoyed the exercise and felt the effectiveness of it. From the last question, the exercise is not always easy for the students. From these results, the students accept the exercise on Monsakun-TapeBlock as enjoyable and useful.

4. Conclusion

Understanding arithmetic word problems can be said as a structural understanding of the relationship between linguistic and mathematical representations. To facilitate learners to build this understanding, this study design and developed a learning environment in which graphic representation bridges the linguistic and the mathematical representation with the quantity roles. The goal of this study is to make learners understand the conditions of addition or subtraction word problem with problem-posing and reflections on the posed problems. For the goal, we propose the task of the conversion among linguistic, mathematical, and graphic representations after problem-posing and design and development a learning environment where learners can conduct the exercise of the conversions. The effectiveness of these exercises is suggested through the experimental use in a public elementary school.

Future tasks will be to verify the difference in learning effect depending on whether or not there is a quantity role matching exercise. In addition to the effects of learning to appear, it is necessary to confirm whether the learner's thought intended in this study appears as the cause.

References

Cummins, D., Kintsch, W., Reusser, K., Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20*, 439-462.

English, L. D. (2003). Problem posing in elementary curriculum In F. Lester & R. Charles (Eds.), Teaching mathematics through problem-solving. Virginia: National Council of Teachers of Mathematics.

Hegarty, M., Mayer, R.E., Monk, C.A. (1995). Comprehension of arithmetic word problems: a comparison of successful and unsuccessful problem solvers. *Journal of Educational Psychology, 87*(1), 18-32.

Heffernan, N., Koedinger, K. (1998). A developmental model for algebra symbolization: The results of a difficulty factors assessment. Proceedings of the twentieth annual conference of the cognitive science society, 484-489.

Hirashima, T., Yokoyama, T., Okamoto, M., Takeuchi, A. (2007). Learning by problem-posing as sentence-integration and experimental use. AIED 2007, 254-261.

Hirashima, T., Yokoyama, T., Okamoto, M., Takeuchi, A. (2008). Long-term Use of Learning Environment for Problem-Posing in Arithmetical Word Problems. ICCE2008, pp.817-824.

Hirashima, T., Kurayama, M. (2011). Learning by problem-posing for reverse-thinking problems. AIED2011, pp.123-130.

Hirashima, T., Yamamoto, S., Hayashi, Y., (2014). Triplet structure model of arithmetical word problems for learning by problem-posing. Human Interface and the Management of Information. Information and Knowledge in Applications and Services, Volume 8522 of the series Lecture Notes in Computer Science, 42-50.

Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *Journal of the Learning Sciences, 13*(2), 129–164.

Mayer, R.E., Thinking, problem solving, cognition, Second ed., pp.455–489, W.H. Freeman, New York, 1992.

Nathan, M., Kintsch, W., Young, E. (1992). A theory of Algebra-Word-Problem Comprehension and Its Implications for the Design of Learning Environments. *Cognition and Instruction, 9*(4), 329-389.

Pólya, G. (1945). How to Solve It. Princeton University Press.

Riley, M. S, Greeno, J. G, Heller, J. I. (1983). Development of children's problem solving ability in arithmetic, H. P. Ginsburg (Ed.). The development of mathematical thinking. New York: Academic Press, 153－196.

Reusser, K. (1996). From Cognitive Modeling to the Design of Pedagogical Tools. In S. Vosniadou et al. (Eds.): International Perspective on the Design of Technology-Supported Learning Environments, LEA, 81-103.

Resnick, L.B. (1983). A developmental theory of number understanding. In Ginsburg, H.P. (Ed.), The development of mathematical thinking (pp. 109-151). New York: Academic Press,

Silver, E. A., & Cai, J. (1996). An analysis of arithmetic problem posing by middle school students. *Journal of Research in Mathematics Education, 27*(5), 521–539.

Willis, G.B. & Fuson, K.C. (1988). Teaching children to use schematic drawings to solve addition and subtraction word problems. *Journal of Educational Psychology, 80*(2), 192– 201.

Reflection Support Environment for Creative Discussion Based on Document Semantics and Multimodal Information

Atsuya SHONO, Yuki HAYASHI* & Kazuhisa SETA
Graduate School of Humanities and Sustainable System Sciences, Osaka Prefecture University, Japan
*hayashi@kis.osakafu-u.ac.jp

Abstract: The purpose of this paper is to realize a mechanism of system processing that stimulates the reflection on a series of research activities with creative discussions (research meetings) as a starting point. In this study, as a clue for systems to capture a part of the semantic contents of the discussion interactions, we utilize *document semantics*, which represent the intentions and contents of the discussion materials, and *multimodal information* exchanged among discussion participants. We propose declarative rules for detecting discussion sections and generating advice that follows the discussion interaction context as a stimulus for reflection. Then, we develop a reflection support environment consisting of a rule creation support system and a reflection support system that embodies the proposed mechanism.

Keywords: Reflection support, multiparty multimodal interaction, document semantics

1. Introduction

An academic research meeting is a place where students belonging to a laboratory can share their research through creative discussions, and is a promising opportunity to cultivate thinking skills (Mori, Hayashi, & Seta, 2019). In general, the proposers (learners) examine the contents of proposals from a multilateral viewpoint (internal conversation), then summarizes the results as discussion materials before the meeting. After the discussion, it is essential that they should reflect on the discussion and examine the inadequate points raised in the meeting. In this reflection, it is desirable not merely to reflect on discussion topics, but also to focus on thinking activities at the time of preparation before the discussion. Such reflection activities contribute to cultivating important perspectives in achieving successive knowledge co-creative discussions, such as a refinement of the internal conversation toward the next research meeting and obtaining perspectives for correctly communicating their intentions to the discussion participants.

This study tackles a research question on how to realize a mechanism of system processing to stimulate the reflection on a series of research activities with such research meetings as a starting point. To approach this, we consider capturing a learners' reflection section by utilizing *document semantics*, which represent the intentions and contents of the discussion materials (the fruits of internal conversation), and *multimodal interaction information* (such as gazing and utterance information) during discussion. In this paper, we propose a novel reflection support environment composed of two systems: a rule creation support system that detects a reflection section aimed at an advisory presentation based on document semantics and interaction information and a reflection support system that applies the created rules and prompts reflection activity.

2. Functional Requirements of Reflection Support Environment

2.1 Requirement Definition

In this section, the functional requirements for the mechanism of the system processing required for our target reflection support environment are listed.

R1. *Mechanism for capturing the context of discussion interactions in a computer-readable format*: Research meetings generally share discussion materials that correspond to the outcome of the learner's preparation activity, and discussion progresses while confirming the same. To provide advice that follows the discussion interaction context as a stimulus for reflection, it is desirable that a system can grasp the semantic content (research content) of the content described in discussion materials as well as what kind of intention (logic composition intention) this is an attempt to explain.
R2. *Mechanism to detect the reflection section by declarative rules with guaranteed reusability*: To adaptively detect the discussion section contributing to the reflection, it is necessary to consider a mechanism with high reusability that does not depend on the session composed of specific discussion participants. Furthermore, by allowing to explicitly associate the defined declarative rules with each layer of the hierarchical interpretation model (Section 2.2), it is desirable that when a new rule is created in any layer, the existence of data detected by such upper or lower layer can be distinguished to create a rule. Here, it is desirable that the processing procedure for detection not be a form embedded in a program but a declarative form in which the intent is easy to understand.
R3. *Mechanism for providing advice according to the focused section of the reflection*: In the discussion reflection, not only is it possible to give the situation at the time of discussion according to the learner's focused section (reflection section detected by the rule satisfying R1 and R2), it is also necessary to be able to provide an introspective advisory toward the thinking activity at the time of preparation.

We focus on document semantics (Section 3.1) and multimodal information (Section 3.2) for R1, and consider a framework of rule creation support to satisfy R2 (Section 3.3). A rule creation support system equipped with this mechanism is then developed (Section 4.1), and a reflection support system is proposed to satisfy R3 (Section 4.2).

2.2 Hierarchical Interpretation Model Based on Interaction Corpus

Hierarchical interpretive model of the interaction (Sumi, Yano, & Nishida, 2010) is a conceptual approach that raises the interpretation from the data belonging to the low-order hierarchy to the high-order hierarchy during a multiparty multimodal interaction. The model consists of four layers: *raw data layer,* which includes simple data sequences such as eye coordinates and voice waveforms; *interaction primitive layer* that corresponds to interaction elements of individual participants such as those who are speaking or looking at someone; *interaction event layer* that combines interaction elements and interprets them such as a joint attention and a mutual gaze; and *interaction context layer* that builds to a higher-order interpretation of interactions regarding the conversational context.

Based on this model, this study considers capturing the discussion sections in the research meeting by combining and interpreting the document semantics and multimodal information of the discussion participants and provides the learners with these sections as reflection points.

2.3 Measurement Environment for Multimodal Information

This study utilizes a multimodal interaction-aware platform for collaborative learning that has a mechanism to capture several verbal and non-verbal information exchanged in multiparty interactions (Sugimoto, Hayashi, & Seta, 2020). The platform is configured to incorporate several learning support tools. Within the system, each participant's video image, utterance timing, interval section of gaze information (gazing at other participants / gazing at material parts) by setting gaze target regions corresponding to each participant's video image and material contents, and so on can be measured using sensing devices. These data are stored in the database for each discussion session, corresponding to the raw data and interaction primitive layers of the hierarchical interpretation model.

3. Rule Creation Based on Document Semantics and Multimodal Information

3.1 Document Semantics

Document semantics are computer-readable datasets that indicate the semantic content of areas on the discussion material. This allows the system to interpret information with the meaning, such as a learner is gazing at the 'research purpose' area from the eye coordinates on the corresponding discussion material area. In this study, we consider two types of semantic information, i.e., *research content semantics* representing the content of the research activity of the learner and *intention semantics* expressing the intention of the logical composition incorporated into the discussion materials.

Research Content Semantics: We use the research activity ontology proposed by Mori, Hayashi, & Seta (2019). In this ontology, the thinking activity required to conduct the ITS research and the requirements of the activity (sub-activities, inputs, and outputs) are structured in a way to take into account the general and specific nature of the research area. Each sub-activity is necessary for achieving another activity. We assume that there are discussion materials in which concepts of the research activity ontology (research content and its linkage) are corresponded by a learner who intends to sufficiently carry out an internal conversation referring to this ontology before the discussion.

Intention Semantics: We utilize the intent ontology of the logical composition about discussion materials, which is an extended ontology proposed by Matsuoka, Seta, & Hayashi (2019). In this ontology, the concept for clearly expressing the logical role of the research content is defined from the viewpoint of the discussion purpose.

In addition to the research content semantics, discussion materials given these intentions reflect the planning activities of the learners who try to design the discussion. Therefore, giving these semantics by learners themselves to the discussion materials as tasks before discussion not only increases the readability of a computer, it also enhances the readiness of learners, and becomes a significant activity from the viewpoint of enhancing the quality of the discussion itself.

3.2 Multimodal Information

In complex communications consisting of several persons, such as discussions, multimodal information such as utterance timing, back-channeling, and gazing information plays a crucial role in addition to the verbal content to be transmitted (Burgoon et al., 2017; Thiran, Marques, & Bourlard, 2009).

This study focuses on two types of modalities that can be detected using the multimodal interaction-aware platform: "utterance information," which is crucial in advancing the discussion, and "gaze information," which represents a part of the thinking of the activity subjects. These communication signals, exchanged in the discussions, assume that they can be treated as having role-based general-purpose and highly reusable interaction information that does not rely on a particular participant by giving the roles of actors such as "instructors" and "proposers."

3.3 Framework for Rule Creation Support System

Based on the multimodal interaction information toward discussion materials given the research content and intent semantics, we propose three types of declarative rules to realize a mechanism for detecting a reflection section that can be applied to sessions consisting of various discussion participants (R2) and a mechanism for generating advice on thinking activities at the time of preparation for discussion (R3).

Initialization Rule: {type, subject, target, rate, inequality, time} · · · (1)

This rule provides the initial settings for the interaction data captured by the multimodal information measurement environment. type is a parameter that specifies target signal out of *Speaking* for handling utterance information, *GazingAtUser* for handling gaze information of the other participants, and a *GazingAtObject* for handling gaze information of the discussion materials. subject is a parameter that specifies the role information of the activity subject (participant) set as a type parameter. If the type is *GazingAtUser*/*GazingAtObject*, the object of the action can be specified as a target parameter from the role information/document semantics. time parameter indicates the arbitrary time interval (e.g., 1, 5, or 10s). This rule detects the interval sections if the section includes the information specified by type, subject, and target more / less than a specific rate, where inequality (more/less) and rate (%) parameters are specified for the detection. The detected discussion sections are stored in the working memory as data corresponding to the interaction primitive layer.

Integration Rule: {layer, function, [arg], [constraint]} · · · (2)

This is a rule for accumulating interpretations in correspondence with each layer of the hierarchical interpretation model based on the data detected by the initialization rule.

layer is a parameter that sets which of the four layers of the hierarchical interpretation model corresponds to the interaction interpretation of the detected data section from this rule. With a function, it is possible to set a function that considers the temporal relationship of the discussion interval (e.g., *Overlap(P1, P2, ...)*, *All(P1, P2, ...)*, *Before(P1, P2)*, *After(P1, P2)*, *Switch(P1, P2)*). Without going into detail in this paper, the arguments of this function are specified in [arg] (e.g., *P1* as arg[0] and *P2* as arg[1]). The function is executed if the data intervals specified in other rules exist in the working memory. [constraint] is a parameter for giving detailed constraints on the role information of the subject/object for the data intervals specified in [arg]. In the integration rule, the data intervals are detected based on the forward chaining method.

Advive Generation Rule: {target_section, feedback} ・・・ (3)
This rule is used for generating advice considering the discussion interval detected by the integration rule as a reflection section. target_section indicates the data intervals (stored in working memory) detected by the integration rule as the reflection interval. In feedback parameter, the rule designer can set the advice that should be checked during this interval in a template format that mixes natural language with specified role information and document semantics (see Section 4.1).

4. Reflection Support Environment

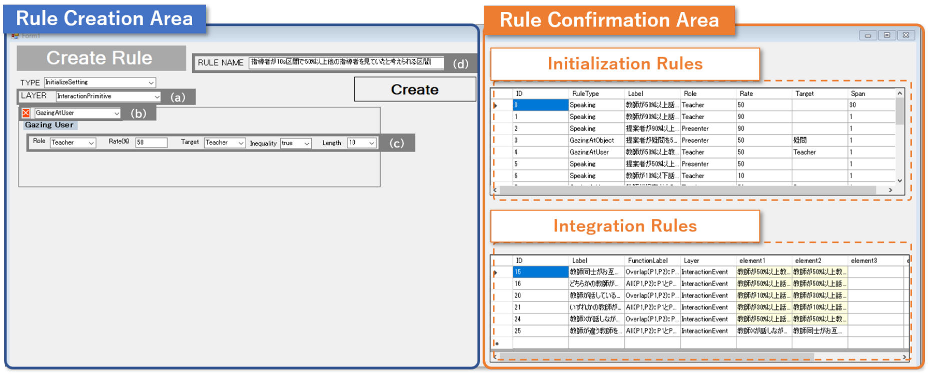
4.1 Rule Creation Support System

Figure 1 shows the rule creation support system. This system consists of a rule creation area and a rule confirmation area. This system has a function to create the three rules explained in Section 3.3.
Function of initialization rule creation (the left side of Fig. 1): This function is used to specify the initialization rule, assuming that the rule corresponds to the interaction primitive layer of the hierarchical interpretation model (Fig. 1(a)). The type parameter for the target interaction data can be specified in the area of Fig. 1(b). Based on the set type, other parameters (i.e., subject, target, rate, inequality, and time) can be specified in the area of Fig. 1(c), respectively. The configured rule is displayed rule confirmation area along with the rule name entered in Fig. 1(d).
Function of integration rule creation (Fig. 2(i)): This function regulates the integration rule. The layer of the hierarchical interpretation model corresponding to the rule can be selected in Fig. 2(i-a). By specifying the function to be applied in Fig. 2(i-b), the corresponding parameters of the function (i.e., [arg] and its [constraint] if necessary) can be set in this area. Figure 3 is the situation where the function is set to 'Overlap,' which detects the intersection of arbitrary intervals. The set rule is listed at the bottom of the rule confirmation area along with the rule name entered in Fig. 2(i-d).

Here, for example, we can set up the initialization and integration rules to capture "*All participants became silent after proposer explains the experimental purpose* (integration rule (7))" by stacking the interpretation as shown in Table 1.

Table 1. *Example of Initialization Rules and Integration Rules*

Example of Initialization Rules	
(1)	{type='Speaking', subject='Proposer', target=null, rate=40(%), inequality='more', time=10(s)} => "*Proposer is speaking*"
(2)	{type='GazingAtObject', subject='Proposer', target='Experimental objectives', rate=60(%), inequality='more', time=10(s)} => "*Proposer is gazing at experimental purpose area in discussion materials*"
(3)	{type='Speaking', subject='Proposer', target=null, rate=10(%), inequality='less', time=10(s)} => "*Proposer is not speaking*"
(4)	{type='Speaking', subject='Teacher', target=null, rate=10(%), inequality='less', time=10(s)} => "*Teacher is not speaking*"
Example of Integration Rules	
(5)	{layer='Interaction-Event', function='Overlap', arg=["*Proposer is speaking*", "*Proposer is gazing at experimental purpose area in discussion materials*"], constraint=[arg[0].subject='Proposer(X)', arg[1].subject='Proposer(X)']} => "*Proposer is explaining experimental purpose*"
(6)	{layer='Interaction-Event', function='Overlap', arg=["*Proposer is not speaking*", "*Teacher is not speaking*"], constraint=[arg[0].subject='Everyone', arg[1].subject='Everyone']} => "*All participants are silent*"
(7)	{layer='Interaction-Event', function='Switch', arg=["*Proposer is explaining experimental purpose*", "*All participants are silent*"], constraint=null} => "*All participants became silent after proposer explains the experimental purpose*"



RULE NAME 指導者が10s区間で50%以上他の指導者を見ていたと考えられる区間

TYPE InitializeSetting

LAYER InteractionPrimitive

GazingAtUser

Role Teacher Rate(%) 50 Target Teacher Inequality true Length 10

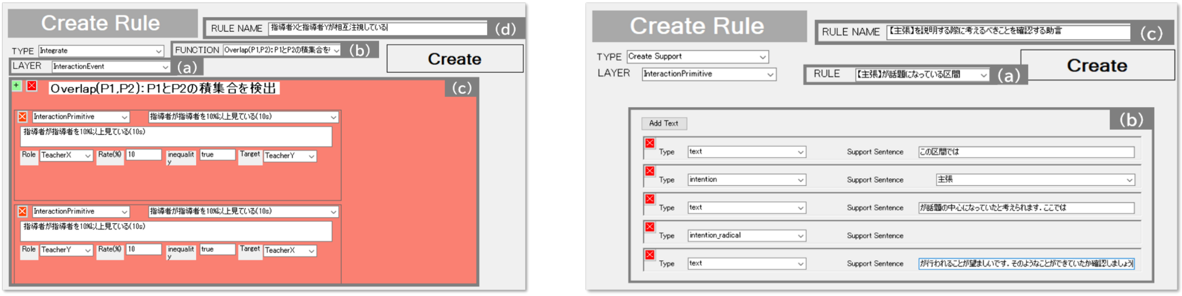
ID	RuleType	Label	Role	Rate	Target	Span
0	Speaking	教師が50%以上話…	Teacher	50		30
1	Speaking	教師が90%以上話…	Teacher	90		1
2	Speaking	提案者が90%以上…	Presenter	90		1
3	GazingAtObject	提案者が疑問を5…	Presenter	50	疑問	1
4	GazingAtUser	教師が50%以上教…	Teacher	50	Teacher	1
5	Speaking	提案者が50%以上…	Presenter	50		1
6	Speaking	教師が10%以下話…	Teacher	10		1

ID	Label	FunctionLabel	Layer	element1	element2	element3
15	教師同士がお互…	Overlap(P1,P2): P…	InteractionEvent	教師が50%以上教…	教師が50%以上教…	
16	どちらかの教師が…	All(P1,P2): P1とP…	InteractionEvent	教師が50%以上話…	教師が50%以上話…	
20	教師が話している…	Overlap(P1,P2): P…	InteractionEvent	教師が10%以上話…	教師が30%以上話…	
21	いずれかの教師が…	All(P1,P2): P1とP…	InteractionEvent	教師が30%以上話…	教師が10%以上話…	
24	教師Xが話しなが…	Overlap(P1,P2): P…	InteractionEvent	教師が50%以上話…	教師が50%以上教…	
25	教師が違う教師を…	All(P1,P2): P1とP…	InteractionEvent	教師Xが話しなが…	教師同士がお互…	

Figure 1. Rule Creation Support System.

Function of advice generation rule creation (Fig. 2(ii)): This is a function used to interpret the detected discussion interval as reflection sections by selecting a rule name as target and setting the concrete advice as feedback. For the target data, the rule name specified in the integration rule can be selected from the drop-down list in Fig. 2(ii-a). The feedback can be set in the area shown in Fig. 2(ii-b). In this area, the rule designer can specify the advice as an array of template formats, including free description text, role information, document semantics, and rational relationships of document semantics (e.g., sub-activity of the research content semantics and consistency of the logical composition of the intent semantics). Here, as advice corresponding to the data interval detected by the integration rule described above ("*All participants became silent after proposer explains the experimental purpose*"), it is possible to create templates as shown below as an example, in order to encourage reflection focusing on the thinking activity in the timing of the preparation for the discussion.

- In this discussion section, all participants became silent after you (Proposer) explained ["experimental purpose" (*concept of research content semantics*)]. In ["proposal" (*concept of intent semantics*)] of ["experimental purpose" (*concept of research content semantics*)], it is desirable to examine ["thinking about experimental subjects", "thinking about evaluation methods" (*sub-activity of "experimental purpose" in research content semantics*)] in advance.
- It is also desirable to examine ["thinking about the validity of proposals and assumptions" (*concept of intent semantics*)] before the discussion. Let's reflect on these points whether you examined them enough before the discussion.



RULE NAME 指導者Xと指導者Yが相互注視している

TYPE Integrate

FUNCTION Overlap(P1,P2): P1とP2の積集合を

LAYER InteractionEvent

Overlap(P1,P2): P1とP2の積集合を検出

InteractionPrimitive 指導者が指導者を10%以上見ている(10s)

指導者が指導者を10%以上見ている(10s)

Role TeacherX Rate(%) 10 inequality true Target TeacherY

InteractionPrimitive 指導者が指導者を10%以上見ている(10s)

指導者が指導者を10%以上見ている(10s)

Role TeacherY Rate(%) 10 inequality true Target TeacherX

(i) Integration rule

RULE NAME 【主張】を説明する際に考えるべきことを確認する助言

TYPE Create Support

LAYER InteractionPrimitive

RULE 【主張】が話題になっている区間

Type	Support Sentence
text	この区間では
intention	主張
text	が話題の中心になっていたと考えられます．ここでは
intention_radical	
text	が行われることが望ましいです．そのようなことができていたか確認しましょう

(ii) Advice generation rule

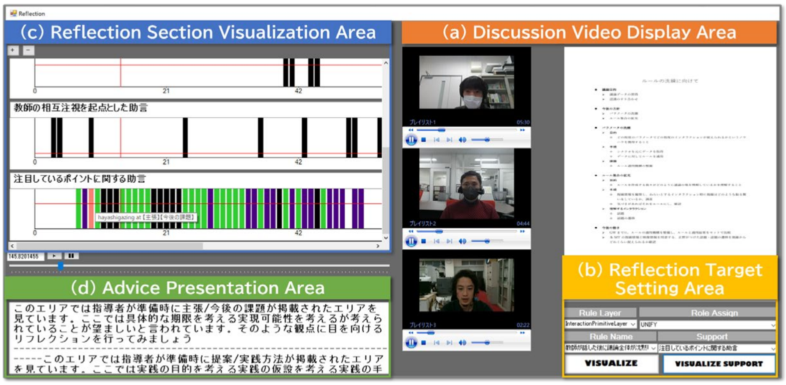
Figure 2. Integration Rule and Advice Generation Rule Creation Area.

4.2 *Reflection Support System*

Figure 3 represents the interface of the developed reflection support system. Before using the system, the user (learner) needs to select a target discussion session for reflection and assign the role information of each participant.

This system has a basic video reflection function that can confirm the synchronized participants' videos and the discussion materials used in the discussion. The user can check the discussion from a given time by operating the seek bar (Fig. 3(a)). In addition, the system has the following two characteristic functions that satisfy the mechanism for providing advice according to the focused reflection section (R3).

Detection function of the target reflection section: Based on the interaction data corresponding to the target discussion session and the role information assigned to the participants, the system applies the initialization and integration rules defined in the rule creation support system at the startup. Then, it


Reflection
(c) Reflection Section Visualization Area
(a) Discussion Video Display Area
(b) Reflection Target Setting Area
(d) Advice Presentation Area
教師の相互注視を起点とした助言
注目しているポイントに関する助言
hayashigazing at【主張】【今後の課題】
0
21
42
145.8201455
VISUALIZE
VISUALIZE SUPPORT

Rule Layer: InteractionPrimitiveLayer
Role Assign: UNIFY
Rule Name: 教師が話した後に議論全体が沈黙!
Support: 注目しているポイントに関する助言

このエリアでは指導者が準備時に主張/今後の課題が掲載されたエリアを見ています。ここでは具体的な期限を考える実現可能性を考えるが考えられていることが望ましいと言われています。そのような観点に目を向けるリフレクションを行ってみましょう

-----このエリアでは指導者が準備時に提案/実践方法が掲載されたエリアを見ています。ここでは実践の目的を考える実践の仮設を考える実践の手

Figure 3. Reflection Support System.

stores the results of detected discussion intervals in the working memory. When there is a result that matches target_section of the advice generation rule, the system generates the instance of feedback of the corresponding rule based on the set template. The names of advice generation rules detected in this way can be available for selection in the area shown in Fig. 3(b).

Visualization function of the reflection target section: When a learner selects the names of advice generation rules as a reflection target section in Fig. 3(b), the detected target sections are displayed in the visualization area in a chart format alongside the discussion time series (Fig. 3(c)). The learner can check the details of the interaction occurring in the arbitrary target section by mouse over operation and can confirm the discussion video from that point by clicking on it. In addition, the advice corresponding to associated with clicked reflection target is displayed in the advice presentation area (Fig. 3(d)).

In this way, the system provides functions that allow learners to concentrate on reflection activities based on the advice, which follows the contents of discussion materials and discussion interaction context, encouraging reflection of thinking activities at the time of preparation.

5. Conclusions

In this study, we discussed a mechanism of system processing that provides advice that prompts a reflection on the discussion context of the research content, including thinking activities at the time of preparation for the creative discussion. At this initial stage, we have confirmed that the proposed system works properly by using experimental data from several sessions. Therefore, future tasks include verifying the effectiveness of the reflection support in the context of authentic research activities.

References

Burgoon, J. K., Magnenat-Thalmann, N., Pantic, M., & Vinciarelli, A. (Eds.). (2017). *Social signal processing*, Cambridge University Press.

Matsuoka, T., Seta, K., & Hayashi, Y. (2019). Internal self-conversation support system by iteration on reflective thinking and research documentation, *Procedia Computer Science*, Vol. 159, pp. 2102–2109.

Mori, N., Hayashi, Y., & Seta, K. (2019). Ontology-based thought organization support system to prompt readiness of intention sharing and its long-term practice. *The Journal of Information and Systems in Education*, Vol. 18, No. 1, pp. 27–39.

Sugimoto, A., Hayashi, Y., & Seta K. (2020). Multimodal interaction-aware integrated platform for CSCL. *Proc. of 22nd International Conference on Human-Computer Interaction*, Vol. 12185, pp. 264–277.

Sumi, Y., Yano, M., & Nishida, T. (2010). Analysis environment of conversational structure with nonverbal multimodal data. *Proc. of International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, Article No. 44.

Thiran, J. P., Marques, F., & Bourlard, H. (Eds.). (2009). *Multimodal signal processing: Theory and applications for human-computer interaction*. Academic Press.

A Coding Mechanism for Analysis of SRL Processes in an Open-Ended Learning Environment

Rumana PATHAN*, Sahana MURTHY & Ramkumar RAJENDRAN
Indian Institute of Technology Bombay, India
*rumana_pathan@iitb.ac.in

Abstract: Open-Ended Learning Environments (OELE) support learning conceptually rich domains. However, widespread use of such OELE has posed several challenges for novice learners, for example, decision making tasks such as trade-off analysis, negotiation, etc. Since the nature of OELE is non-linear and open-ended, it requires the need of employing several self-regulatory processes such as planning, cognitive strategies, metacognitive monitoring, etc. To analyse these self-regulated learning (SRL) processes in an OELE, we introduce and discuss a coding mechanism based on Pintrich's framework of SRL and the design of a learning environment. The mechanism discusses several cognitive and metacognitive processes and observable indicators that can be representative/suggestive of a specific regulatory process that a learner might be displaying. To test the mechanism, a retrospective think-aloud (N=10) was conducted. Our primary contribution is developing and implementing the proposed coding mechanism. The findings of the work presented in the paper indicate a detailed understanding of the regulatory processes employed by learners while solving an open-ended problem in an OELE.

Keywords: Self-regulated learning, coding mechanism, open-ended learning environment, retrospective think aloud

1. Introduction

Open-ended learning environments (OELE) provide learners with opportunities for inquiry and complex problem-solving by presenting them with authentic contexts, stimulating learning tasks, and tools and resources to explore. Such learning environments stimulate learner abilities to expound decision making tasks like trade-off analysis, critical thinking, negotiation, etc. (Land, 2000). Thus, novice learners are required to employ several self-regulatory processes such as planning, cognitive strategies, monitoring, etc. (Azevedo et al., 2010).

To learn effectively in an OELE, a learner must analyse the learning context, set sub-goals, decide learning strategies to employ, assess the strategy, and monitor emerging understanding (Azevedo et al., 2010). Such learning involves deploying several cognitive, metacognitive, motivational and behavioural processes (Pintrich, 2000). Therefore, understanding these processes has become important for researchers to grasp the complex nature of SRL (Greene & Azevedo, 2009). Existing frameworks to understand SRL have been applied in hyperlink-based environments, in which the primary task is to read, assimilate, or synthesise text using various learning resources. In comparison, OELEs developed for complex problem solving may include tools such as a simulator, causal map builder, etc. Thus, existing frameworks are not sufficient to identify SRL processes in OELE.

To address the above-mentioned challenge, this paper proposes a coding mechanism based on Pintrich's (2000) framework of SRL to identify SRL processes in a OELE developed for problem-solving. This framework is suitable because it classifies different phases (i.e., forethought and planning, monitoring, control, and reaction/reflection) and areas (i.e., cognition, motivation/affect, behaviour, and context) of regulation as a heuristic to organise and understand SRL. To demonstrate the usability of the proposed mechanism, we have conducted a study with ten learners interacting with MEttLE (Modeling-based Estimation Learning Environment), a web-based OELE for estimation problem-solving using retrospective think-aloud (rTA) protocol. The proposed mechanism contains cognitive and metacognitive processes derived from both Pintrich's framework (2000) and the design of MEttLE.

2. Literature Review

SRL is an extraordinary umbrella that considers several aspects that influence learning (e.g., self-efficacy, volition, cognitive strategies) in a holistic approach (Panadero, 2017). Several SRL frameworks, models, and theories explain how cognitive, metacognitive and contextual factors influence the learning. For example, Zimmerman and Schunk (2001) first outlined SRL process by proposing a cyclic model of SRL with three phases, i.e., forethought, performance, and self-reflection. Pintrich (2000) extended this model to include four phases (i.e., forethought and planning, monitoring, control, and reaction/reflection). In 1998, Winne and Hadwin envisaged a cognitive structure that involved variables at the personal level processes at the task and personal level. Hence, different models predicate slightly distinct views on how learners self-regulate.

Whilst several theoretical models of SRL exist, the measurement of SRL remains a central issue in this field of research. An objective way of identifying and measuring student regulation is using online measures such as think-aloud data (Greene & Azevedo, 2009). Similarly, the stimulated recall method, a type of retrospective think-aloud protocol, can be used. In stimulated recall, learners' interaction data (e.g., screen capture video) is played back. They are asked to verbalise what they did at each point in the problem-solving process and reason for their actions. Such verbalisations can elucidate SRL processes and help us understand the dynamic nature of SRL. Recently, trace data, also known as event logs or log data, is used to measure SRL (Siadaty et al., 2016; Munshi & Biswas, 2019). Traces capture learner actions on the fly along with the context. Although trace data has a methodological advantage over think-aloud and other self-reports, it cannot and should not be considered the only method for measuring SRL processes (Winne, 2010).

Several existing learning environments support SRL processes, such as Metatutor (Azevedo et al., 2010) and Learn-B (Siadaty, Gasevic & Hatala, 2016). Although learning environments support SRL processes, very few measure such processes using theoretically grounded frameworks. We identified three such frameworks from the literature, viz. CAMM (Cognitive Affective Metacognition Motivation) model of SRL by Azevedo and colleagues (2010), a framework of self-regulated hypermedia learning by Bannert (2007), and trace-based microanalytic framework by Siadaty and colleagues (2016). While these frameworks to measure SRL processes exist, the tasks associated with the learning environments in question are mostly reading and assimilating. None of the environments supports problem-solving tasks including the use of tools (e.g. simulator). In the following sections, we are motivated to understand SRL processes associated with problem-solving tasks by proposing a coding mechanism based on Pintrich's framework and design of one such OELE, viz. MEttLE.

3. Coding Mechanism to Capture SRL Processes

This section describes the procedure used to combine theory and the pedagogical design of MEttLE to derive the coding mechanism for capturing learners' SRL processes in MEttLE.

3.1 Theoretical Basis

To identify the theoretical underpinning, we reviewed several SRL models, frameworks, and theories synthesised in section 2. We found Pintrich's SRL model (2000) most suited for our task of modelling regulatory processes. The framework describes and classifies several regulation processes to reflect goal setting, monitoring, control, and reaction and reflection regulatory processes across areas of regulation such as cognition, motivation/affect, behaviour, and context. We will be focused on discussing the extent of the area 'regulation of cognition' in the 'context' of MEttLE.

In the area 'regulation of cognition', the 'forethought, planning and activation' phase involves planning, goal setting and activation of relevant knowledge of the task. Similarly, 'monitoring' concerns various monitoring processes representing metacognitive awareness. Likewise, 'control' involves efforts to control and regulate different aspects of the task by selecting and adapting cognitive strategies for thinking and learning. Finally, the 'reaction and reflection' phase represents various reactions and reflections on the task. It is crucial to correctly infer the learners' verbalisation and associate it with the appropriate regulatory process. To do so, we reviewed literature and extracted a list of observable indicators that are representative/suggestive of a specific regulatory process that a learner might be going through. Table 1 outlines regulatory processes described by Pintrich (2000) and their corresponding indicators found in related literature. For example, if a *"learner begins a task by setting specific goals for learning"*, it suggests that the learner is displaying 'target goal setting'.

Table 1. *Regulatory Processes Defined in Pintrich (2000) and Their Corresponding Indicators*

Phases	Regulatory processes		*Description of indicators*
Planning & activation	i.	Target goal setting	*Learner may begin a task by setting specific goals for learning*
			Learner may set specific goals for time use
			Learner may set specific goals for eventual performance
			Adjust or change the goal during task performance
	ii.	Activation of prior content knowledge	*Learner activates prior knowledge by actively searching their memory for relevant prior knowledge*
			Learner can activate prior knowledge in a planful and regulatory manner through prompts and self-questioning activities
			Learner constructs better problem representation
	iii.	Cognitive tasks	*Learner has knowledge how task variations can influence cognition*
			Learner knows that some tasks are more or less difficult
	iv.	Cognitive strategies, declarative	*Learner has knowledge that some strategies can help in learning*
			Learner has knowledge of 'what of cognition'
			Learner has knowledge of different cognitive strategies, such as rehearsal or elaboration, that can be used for learning
	v.	Cognitive strategies, procedural	*Learner knows how to perform and use a cognitive strategy*
			Learner knows that there are different strategies, and how to use
	vi.	Cognitive strategies, conditional	*Learner knows when and why to use a cognitive strategy*
			Learner knows that one strategy may be appropriate in some contexts, and may not in some.
Monitoring	vii.	Judgement of learning and comprehension monitoring	*Learner becomes aware that he does not understand something they just read or heard*
			Learner becomes aware that he understood something
			Learner becomes aware that he is reading too quickly or slowly
			Learner monitors reading comprehension by asking questions
			Learner decides if he is ready to take a test on the material he read
			Learner judges his comprehension of a lecture
			Learner judges whether he could recall the information for a test
	viii.	Feeling of knowing	*Learner cannot recall something when called upon to do so, but knows it*
			Learner cannot recall something when called upon to do so, but have strong feelings that he/she knows it.
			Learner is aware of reading something in the past and having some understanding of it, but not being able to recall it on demand
			Learner recalls teacher discussing in class, but is not able to recall
Control & regulation	ix.	Selection and adaptation of control strategies	*Strategy may include use of imagery to help encode information*
			Strategy may include use of imagery to help visualise correct implementation of a strategy
			E.g., strategies use of mnemonics, paraphrasing, summarising, outlining, networking, constructing tree diagrams, note-taking, etc.

Reaction & reflection	x.	Cognitive judgments	*Learner evaluates his performance*
	xi.	Adaptive attributions	*Learners make attributions to low efforts or poor strategy use*
			Learner make attribution of success to self
			Learner make attribution of success to external factor
			Learner make attribution of failure to self
			Learner make attribution of failure to external factor

3.2 Design of Problem-Solving OELE: Mettle

MEttLE is designed to support novice estimation problem solving (Kothiyal & Murthy, 2018). For instance, learners estimate the electrical power required to design the motor of a racing car, with given specifications such as wheel diameter, distance, etc. MEttLE offers metacognitive prompts, expert guidance and hints to help learners plan, monitor, and reflect. Similarly, it also consists of various tools and resources such as a simulator, calculator, info center, scribble pad, causal map builder, equation builder and a problem map. MEttLE is also capable of logging student data (Pathan et al., 2019). With the help of MEttLE's design, we scoped down the regulatory processes supported in MEttLE by various features and their corresponding indicators (Table 2). For example, 'productive planning' in MEttLE is supported by 'metacognitive prompts' and is indicated by *"learners writing planning statements using planning question prompts".*

Table 2. *Regulatory Processes Supported by the Design of Mettle and Their Corresponding Indicators*

SRL processes	MEttLE context	*Description of indicators*
1. Planning	Metacognitive prompt	*Learner writes planning statements using prompts*
2. Monitoring		*Learner writes monitoring statements using prompts*
3. Model building techniques	Simulator	*Learner uses variable manipulation simulation, with implicit guidance to incorporate problem context*
	Statement builder	*Learner uses fictive motion words from the word bag*
	Causalmap builder	*Learner uses causal mapping tool*
	Equation builder	*Learner uses drag and drop parameters and mathematical relationships relevant to the problem*
4. Estimation reasoning	Question prompt	*Learner uses prompts to do estimation reasoning*
	Hints & guide me	*Learner uses hint /guide me to do estimation reasoning*
5. Gather context specific knowledge	Problem context	*Learner reads information on the problem context*
	Information center	*Learner uses infocenter to gather context specific knowledge*
6. External representation	Scribble pad	*Learner uses scribble pad*
7. Reflection on process	Reflection prompt	*Learner uses prompts to do reflection*
	Problem map	*Learner uses problem map to do reflection on problem solving process*
8. Evaluation during model building	Evaluation prompt	*Learner uses evaluation and contextualisation prompts*

3.3 Merging SRL Processes from Theory and OELE to Develop a Coding Mechanism

Table 3 delineates the merged coding mechanism. To consolidate the regulatory processes found, we aligned each MEttLE specific indicator to its closest mapping found in Pintrich. For example, the indicator from MEttLE specific process 'planning' *("learner writes planning statements using planning question prompts")* is similar to indicator of Pintrich's process 'target goal setting' (*"learner may begin* *a task by setting specific goals for learning”).* Hence, the indicators of ‘productive planning’ and ‘target goal setting’ can be merged in the coding mechanism, as shown in Table 3. Likewise, we arranged all the MEttLE specific regulatory processes under Pintrich’s classification of metacognitive processes.

Table 3. *Coding Mechanism Based on Pintrich and Design of OELE to Capture SRL Processes*

Phase	Regulatory processes	Phase	Regulatory processes
Planning and activation	i. Target goal setting 1. Planning	Control and regulation	ix. Selection and adaptation of control strategies
	ii. Activation of prior content knowledge		3. Model building techniques
	iii. Cognitive tasks		4. Estimation reasoning
	iv. Cognitive strategies, declarative		5. Gather context specific knowledge
	v. Cognitive strategies, procedural		6. External representations
	vi. Cognitive strategies, conditional	Reaction and reflection	x. Cognitive judgements
Monitoring	vii. Judgement of learning and comprehension monitoring 2. Monitoring		xi. Adaptive attributions
			7. Reflection on process
	viii. Feeling of knowing		8. Evaluation during model building

4. Modelling SRL Processes Using the Coding Mechanism in MEttLE

The research goal of this study was to implement the proposed mechanism using rTA verbalisations of learners interacting with MEttLE and thus validate the existence of the indicators. Ten learners (6 male, 4 female) who had completed at least one year in Engineering participated in the study. The study was conducted in a lab set-up; wherein individual learners solved a complex engineering problem in MEttLE. within 60-90 minutes.

Two researchers independently coded 40% of learner data collected from 10 learners’ interviews to establish the reliability of our coding mechanism. The researchers interpreted each phrase with the help of indicators provided in Tables 1 and 2, and used the closest one to code the phrase. For example, the learner statement *“I learned the concept of friction that if a body is moving at constant speed, the floor or the track, or on anything that its moving, that thing will oppose it”*, is closely identified with the indicator “Learner activates prior knowledge by actively search their memory for relevant prior knowledge”. Thus, this phrase is coded as ‘Activation of relevant prior content knowledge’. The inter-rater reliability of the coded interviews was calculated as cohen’s kappa 0.83, indicating a strong agreement level. The rest of the interviews were then coded by one of the two researchers. Total 970 phrases were coded, and the rest were marked NA if they did not comply with any indicator.

We analysed the coded interviews to examine and understand the occurrence of various regulatory processes and their indicators. The average frequency of total processes per learner was 97. We found all the indicators listed in the proposed mechanism. The most common indicators found were related to ‘model building’ and ‘judgement of learning and comprehension monitoring’. Out of 19 processes described in the mechanism, 14 processes commonly occurred. The SRL processes that occurred scarcely are activation of metacognitive knowledge (declarative, procedural, conditional), productive reflection on the process, and use of external representation. On grouping the indicators under their respective major categories, we found that learners displayed the maximum number of average processes under ‘Control and regulation’ (31%) and planning & activation (31%), followed by monitoring (23%) and reaction and reflection (15%).

5. Discussion and Conclusion

This paper demonstrates developing and implementing a coding mechanism to capture SRL processes in an OELE. The mechanism built on the theoretical basis of Pintrich and the pedagogical design of the OELE classifies SRL processes into four major categories. Although in its preliminary design stage, such a mechanism is developed because no existing frameworks are specifically designed to capture SRL processes in an OELE to solve complex problems. To demonstrate the use, we implemented the mechanism on verbalisations produced by ten learners.

We found 31% of regulatory processes associated with regulation of control, i.e., employing several learning strategies. Similarly, in a study conducted by Azevedo et al. (2010), learners think-aloud data indicated that learning strategies were deployed most often. In the control phase, indicators associated with model building strategies, such as simulation, causal map builder, etc., are employed most times. These indicators are particularly associated with tools found in complex problem-solving OELE and are hence essential to capture. While frequency analysis using rTA was useful, it did not capture time-sensitive information, such as the relationship between two processes. Thus, devoiding us opportunities to apply relationship mining algorithms such as sequential pattern/process mining. Hence, to capture temporal data, we plan to collect concurrent think-aloud data in the future.

While the existing coding mechanism is designed for MEttLE, the procedure to extract and merge the processes and indicators remains generic for OELE's designed to solve complex problems. We believe it is generalisable because our procedure ensures the inclusion of observable behaviour using both theory and design of the OELE. To implement the coding mechanism in another OELE, the following steps will have to be ensured, 1) identify features/affordances in the new OELE, and the SRL processes it braces, 2) extract a list of observable indicators that suggest the use of SRL processes supported by OELE, 3) find the closest alignment between OELE specific indicators and Pintrich indicators outlined in Table 1, and 4) merge the OELE specific SRL processes with the processes described in Pintrich. In future, we plan to conduct think-aloud studies with a larger sample to validate our coding mechanism. The resulting time-sequenced process series can be annotated with corresponding trace data to identify learner SRL processes automatically in an unobtrusive manner.

Acknowledgements

We would like to thank Dr. Aditi Kothiyal for her contributions in designing & conducting the study.

References

Azevedo, R., Moos, D. C., Johnson, A. M., & Chauncey, A. D. (2010). Measuring cognitive and metacognitive regulatory processes during hypermedia learning: Issues and challenges. Educational psychologist.

Bannert, M. (2007). Metacognition when learning with hypermedia. Waxmann Verlag.

Greene, J. A., & Azevedo, R. (2009). A macro-level analysis of SRL processes and their relations to the acquisition of a sophisticated mental model of a complex system. Contemporary Educational Psychology, 34(1), 18-29.

Kothiyal, A., & Murthy, S. (2018). MEttLE: a modelling-based learning environment for undergraduate engineering estimation problem solving. Research and practice in technology enhanced learning.

Land, S. M. (2000). Cognitive requirements for learning with open-ended learning environments. Educational Technology Research and Development, 48(3), 61-78.

Munshi, A., & Biswas, G. (2019, June). Personalisation in OELEs: developing a data-driven framework to model and scaffold SRL processes. In International Conference on Artificial Intelligence in Education

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. Frontiers in psychology, 8, 422.

Pathan, R., Shaikh, U., & Rajendran, R. (2019, December). Capturing learner interaction in computer-based learning environment: design and application. In 2019 IEEE Tenth International Conference on Technology for Education (T4E) (pp. 146-153). IEEE.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In Handbook of self-regulation (pp. 451-502). Academic Press.

Siadaty, M., Gasevic, D., & Hatala, M. (2016). Trace-based micro-analytic measurement of self-regulated learning processes. Journal of Learning Analytics, 3(1), 183-214.

Winne, P. H. (2010). Improving measurement of self-regulated learning. Educational Psychologist.

Winne, P. H., & Hadwin, A. E. (1998). Studying as self-regulated learning (pp. 291-318). Routledge.

Zimmerman, B. J., & Schunk, D. H. (Eds.). (2001). Self-regulated learning and academic achievement: Theoretical perspectives. Routledge.

Viewpoint Transformation Training System Based on Discovery of Relationships between Objects

Kota KUNORI [a*] **& Tomoko KOJIRI**[b]
[a]*Graduate School of Science and Engineering, Kansai University, Japan*
[b]*Faculty of Engineering Science, Kansai University, Japan*
[*]k194388@ kansai-u.ac.jp

Abstract: Although having multiple viewpoints is important for utilizing objects in various ways, some people can only focus on limited aspects of objects. An object's viewpoint is regarded as its role, as seen from its relationships with other objects. To discover new roles by finding relationships with other objects, another object must be recognized. Thus, our study proposes a method of two steps that transforms viewpoints consisting of "finding other objects in a space (object discovery step)" and "considering roles based on the relationships between a target object and other objects (role discovery step)." In addition, our study constructs a viewpoint transformation training system to acquire multiple viewpoints in the mathematical world, where objects correspond to individual figures whose viewpoints are their roles viewed from the stances of other figures.

Keywords: Viewpoint transformation, role of objects, relationship between two objects, mathematical diagram

1. Introduction

We are often suggested "to see things from a different viewpoint," for example, when we are worried about something or trying to solve a problem. By seeing things from a different viewpoint, we can sometimes find new solutions. Looking at things from a new viewpoint is also important for exploiting objects. Assume the utilization of a pair of sunglasses. If we only see them as an item that protects our eyes from the sun, we are unable to use them for any other purposes. However, if we frame them as a "fashion item," they become a lifestyle accessory.

Once we see something from one aspect, changing that viewpoint is difficult. The objective of our study is to propose a method that changes viewpoints for recognizing things/objects and develops a training system for it. One way to change viewpoints is using another person's perspective. Han et al. proposed an idea generation support system which acquires other person's perspective from social networking platforms as Twitter and Wikipedia (Han, Park, Forbes & Schaefer, 2020). If a user wants to generate new ideas for a specific thing, he or she can retrieve a huge number of posts from various social networking platforms. Itou et al. proposed an idea generation support system that presents hints for brainstorming associated with a theme (Itou, Higashi & Munemori, 2015). It previously collected web texts related to that theme and classified words into clusters based on the degree of word co-occurrence to make a hint database. Both of these studies use other people's viewpoints to derive new ideas. However, sometimes the viewpoints of others are unavailable when we need to solve problems or derive ideas. To obtain various viewpoints for cases, a method must be acquired with which we can transform our viewpoints by ourselves.

The objective of our study is to propose a method for changing the viewpoint of an object. Things/objects are often recognized by their roles, which are defined in combination with other objects. To change the viewpoint of things/objects is regarded as discovering their roles by identifying meaningful relationships with other objects. Therefore, our study proposes a method consisting of two steps; to find objects and to find the relationships between two objects. In addition, we propose a system with which to experience the proposed method.

Currently, our system limits the problem space to the mathematical world.

2. Method for Changing Viewpoints

The viewpoints of objects are regarded as their roles for other objects. To discover a new role, we must notice not only relationships with other objects but also the objects themselves. Our study proposes a method for discovering viewpoints consisting of two steps: 1. finding objects in a problem space (object discovery step) and 2. defining the role by considering the relationships between the target object and other objects (role discovery step). After recognizing different objects, it is easier to focus on various combinations of objects and ponder their relations.

Note the method using Figure 1, where the existing figures are points A, B, C, and D, line segments AB, AC, BC, AD, and CD, triangles ABC and ACD, and quadrilateral ABCD. In the object discovery step, these figures are recognized by considering the combinations of all the points. In the role discovery step, the relationships of pairs of figures are considered. For example, line segment AC is an adjacent side with line segments AB, BC, AD, and CD as well as one side of a triangle with triangles ABC and ACD.

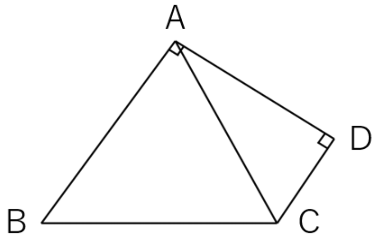


Figure 1. Mathematical Diagram.

3. System for Acquiring Method for Changing Viewpoints

To acquire the steps for changing viewpoints, these steps described in Section 2 must be consciously practiced. Our study develops a system where the steps for changing viewpoints can be experienced individually using mathematical diagrams. The more figures are discovered, the more viewpoints are derived. Therefore, the object discovery step's goal in our system is to find all the figures from a given diagram. In the role discovery step, the relationships of all the discovered figures should be considered.

Figure 2 show the interfaces and the interaction between user and our system. Our system consists of two interfaces that enable the activities of the object discovery step and the role discovery step.

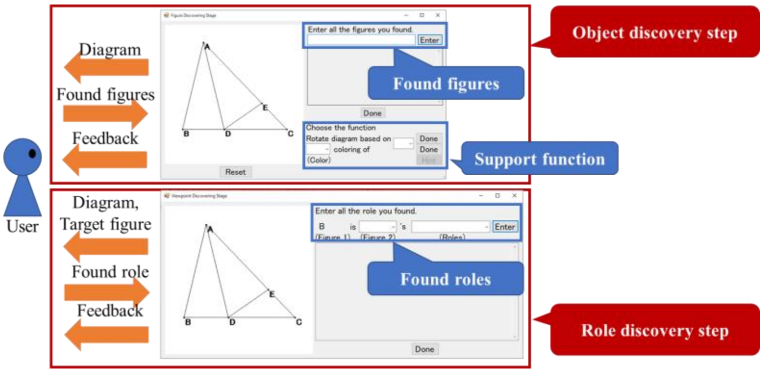
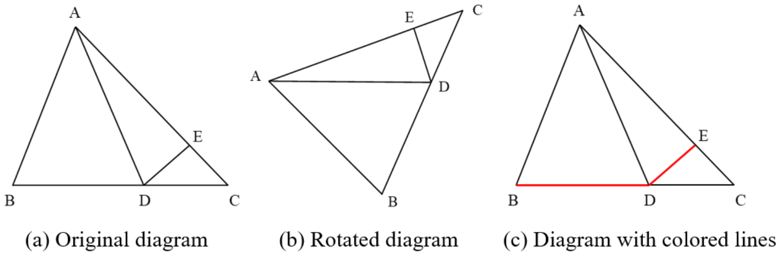


Figure 2. System Interface.

The object discovery interface presents a diagram and lets the user inputs all the figures found in the diagram. It also provides two support functions for finding figures. One function is a rotation function. One of the reasons why we cannot find particular figures is that their orientation is unfamiliar. Especially in the case of polygons, their familiarity is affected by the angle at which the diagrams are drawn. To solve this difficulty, the rotation function enables users to orient the diagram in preferred angle. Figure 3b is the diagram which shows the result of applying the rotation function to Figure 3a.

The other function is a coloring function. Another reason for the inability to detect figures is the presence of other figures. In Gestalt psychology, if a person recognizes some polygons in a diagram, he probably cannot recognize any other polygons (Richard Wiseman, Caroline Watt, Kenneth Gilhooly and George Georgiou, 2011). To recognize other polygons, since they consist of several line segments, different combinations of line segments must be focused on. By emphasizing some line segments, we may find different combinations of line segments with emphasized lines. To encourage users to change the focusing line segments, the coloring function enables users to set preferred colors to line segments. Figure 3c is the diagram which shows the result of applying the coloring function to Figure 3a.



(a) Original diagram (b) Rotated diagram (c) Diagram with colored lines

Figure 3. Example of Applying Support Function**s**.

The object discovery interface has all the figures and their relationships as correct answers. When all the figures have been discovered, the system activates the role discovery interface. If undiscovered figures remain, the system says, “undiscovered figures remain”, and user can use hint function. Hint function generates a hint message for indicating the way of using the support function. For example, if the undiscovered figures are polygons, it gives a message, “Why don't you try coloring the line AD?”

The role discovery interface gives the target figure to focus on for finding the roles and lets the user inputs the discovered roles by finding as many relationships with other figures as possible. It compares the input roles with the correct answer. If the user could not find all relations, the interface shows correct relations one by one as feedback. When all the roles have been displayed as feedback, the interface is returned to the role input mode and a different target figure is given, which allows the user to detect a new role for the given figure.

4. Conclusion

This paper proposed a method for changing viewpoints to discover objects and think about the roles they play in relation to other objects. It also developed a system for training this method in the domain of mathematical diagrams.

If we acquire various viewpoints, we can solve more problems can be solved. Such experience may promote the recognition of the importance of acquiring various viewpoints. Our future work will extend this system for applying developed roles to answer mathematical problems and increase awareness of the importance of changing viewpoints.

References

Han, J., Park, D., Forbes, H. & Schaefer, D. (2020). A Computational Approach for Using Social Networking Platforms to Support Creative Idea Generation. Procedia CIRP, 91, 382-387.

Itou, J., Higashi, T. & Munemori, J. (2015). Proposal and Comparison of an Idea Generation Support System Presenting Words for the Use of Scarce Knowledge People. *Procedia Computer Science*, 60, 918-925.

Wiseman, R., Watt, C., Gilhooly K. and Georgiou, G. (2011). Creativity and Ease of Ambiguous Figural Reversal. *British Journal of Psychology*, 102, 615-622.

Support System for Understanding Intention in Communication Using Diagrams

Koushi UEDA[a*] **& Tomoko KOJIRI**[b]
[a]*Graduate School of Science and Engineering, Kansai University, Japan*
[b]*Faculty of Engineering Science, Kansai University, Japan*
*k581376@kansai-u.ac.jp

Abstract: Readers understand the messages from the diagrams by grasping their features. However, some fail to notice the features and do not correctly grasp the messages. For the first step of reading the messages, this paper focuses on intentions included in diagrams as "emphasize," "classify," and "continuous change" and proposes a method for comprehending intentions from diagrams. To comprehend intentions, the differences in the attributes of figures must first be noticed, and then the intentions indicated by these differences must be considered. Our system encourages users to grasp intentions by following these steps. It also develops a system for experiencing the method.

Keywords: Intention of diagrams, reading diagrams, discovering attributes of figures

1. Introduction

Diagrams are often used to convey messages, because many people find them intuitive and easy to understand (Bondy & Frost, 1993). The creators of diagrams represent the contents and the intentions of messages as the features of figures. The contents of messages are often represented by the shapes of figures that can represent the meaning of the contents. For example, a triangle with a circle on its top often represents a human being. Intentions are represented by the distribution of attribute values. When the creators want to emphasize something, they put a different color on the target figure or increase its size.

On the other hand, readers understand the messages from the features of diagrams. However, some fail to notice the features and fail to correctly grasp the message. Helping such readers learn how to read messages is effective for smooth communication. This research aims to provide a method that readers can use for understanding messages from diagrams. For the first step of attaining to this research goal, this study focuses on intentions, such as "emphasize," "classify," and "continuous change," and proposes a method for reading intentions from diagrams.

As for human-computer interaction, several studies have developed systems that can read diagrams (Swinkels, Claesen, Xiao & Shen, 2018). In these studies, systems obtained the ability to understand the diagrams, but they do not support readers' ability to comprehend the diagrams.

This study proposes a method for comprehending intentions from a diagram and provides an environment for experiencing our method.

2. Method of Comprehending Intention from Diagrams

Readers grasp the intentions of creators by the values of the attributes of figures and their distributions. When the same value is assigned to all figures, there is no significant intention, but when different values are assigned, they suggest intention. The meaning of intentions differs by the distribution of the figures for each value. If the number of values are two and when the number of figures that belong to one value is small and the other is large, the creator may "emphasize" the figures in the smaller group. If the number of values exceeds two, the creator may "classify" the objects by values. When the number of values exceeds two and the attribute is a continuous value, such as size, the creator may indicate the "continuous change" of the features of objects.

An example of inferring intentions is shown with Figure 1. If readers focus on the shapes of the objects, they will divide them into three groups: ovals, rectangles, and arrows. Since the number of group exceeds two and the shapes are not continuous values, the intention is to "classify" the objects by their shapes. If readers only focus on the arrows, they might divide them into two groups by color: blue and red. Since only one object belongs to red, the intention, represented by the color of the arrow, probably "emphasizes" the red object.

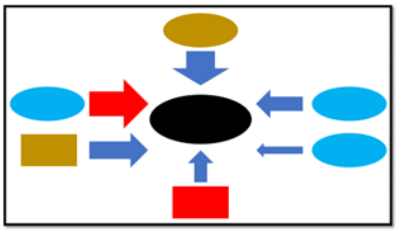


Figure 1. Example of Diagram.

Common and different values of attributes are differentiated by the target sets of figures. For example, in Figure 1 when focusing on all the figures, since no attribute values are common to all the figures, such attributes as color or size suggest intention. On the other hand, when the ovals are focused on, shape and size may be common to all the figures, but their color is different. Perhaps color represents intention. Therefore, to find attributes that may indicate intention, the focus group of figures need to be changed according to their common attributes.

This study defines the following method of comprehending intentions from a diagram:

Step 1: determine a group of focus figures, which is called a target group.

Step 2: select one attribute as a focus attribute.

Step 3: divide the figures in the target group into new groups based on the values of the focus attribute.

Step 4: consider the intention based on the number of figures for each new group.

Step 5: set the states of the target group as verified and return to step 1 until the states of all the groups are verified.

The current target intentions of this research are "emphasize," "classify," and "continuous change."

3. System for Obtaining Method of Reading Intention in Diagrams

This research proposes a system in which users can experience the reading intentions of diagrams with a defined method. Figure 2 shows its system configuration. The diagram database contains diagrams to provide to users and attribute values that represent intentions as correct answers.

The system has four interfaces. The group selection interface corresponds to Step 1 of the method (Figure 3). It gives a user a diagram in a diagram display area from the diagram database. In the interface, the user selects a target group comprised of figures. If users are able to find an attribute in the target group that represents intentions and push Next button, the attribute selection interface corresponding to Step 2 is shown for inputting the target attribute. If they are not able to find the attribute, the attribute discovery support functions presents in the hint area provides hints.

The attribute discovery support function provides two function for supporting users who cannot identify which attributes to focus on: the attribute unification function and the group concentration function. One reason that users cannot discover particular attributes is that they are overly focused on the differences in the values of already discovered attributes. If there are no differences in the values of such attributes, they will not concentrate on them and will look for other attributes. Therefore, attribute unification function changes the values of the already found attributes of the figures to one unified value. Another reason that users cannot discover certain attributes is that the diagram has too many figures. If it only shows figures whose attributes need to be discovered, users may more easily find the attributes. Group concentration function eliminates the figures other than those whose attributes should be discovered.

After selecting the attribute in the attribute selection interface, the group generation interface corresponding to Step 3 is invoked (Figure 4). In this interface, users can divide the figures in the target group by the values of the focus attributes and generate new groups for each value of the focus attributes. By pushing Next button, the intention and meaning input interface corresponds to Step 4 appears and allows users to identify the types of intentions and assign meanings to target groups. When the assignment is completed, the group selection interface is activated again. By comparing the user’s input and the correct answer, the interface judges whether all the intentions of the correct answer were derived; if not, the system starts the method from the beginning.

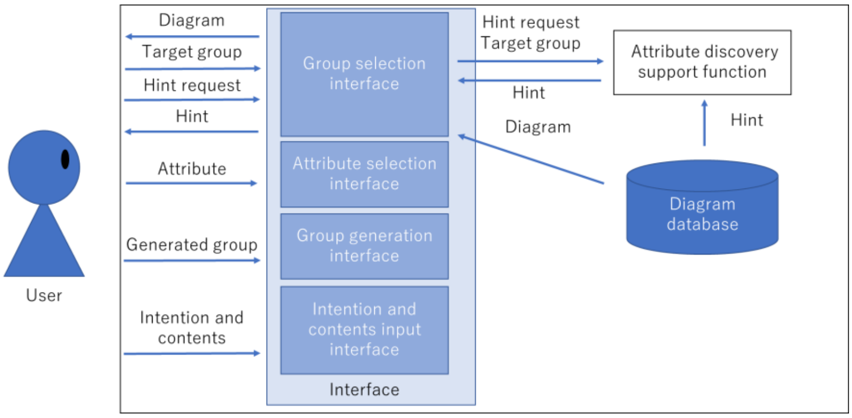
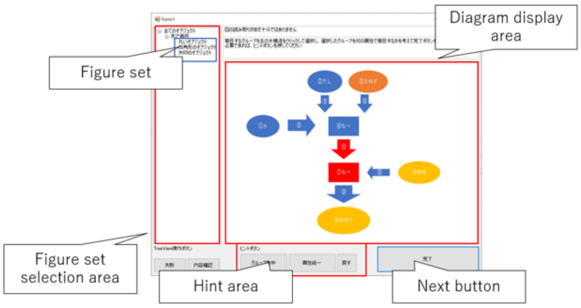
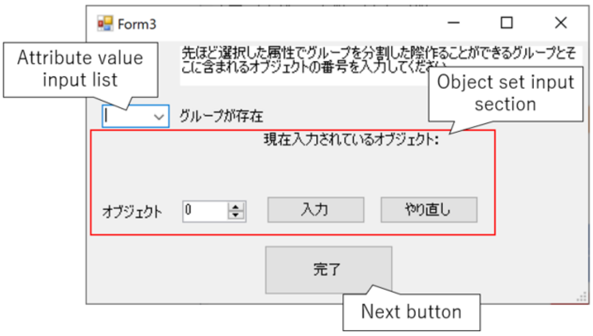


Diagram
Target group
Hint request
Hint
Attribute
Generated group
Intention and contents
User
Group selection interface
Attribute selection interface
Group generation interface
Intention and contents input interface
Interface
Hint request
Target group
Hint
Diagram
Attribute discovery support function
Hint
Diagram database

Figure 2. System Configuration.





先ほど選択した属性でグループを分割した際作ることができるグループとそこに含まれるオブジェクトの番号を入力してくださ

Figure 3. Group Selection Interface.

Figure 4. Group Generation Interface.

4. Conclusion

This paper proposed a method for understanding the intentions drawn in diagrams. It consists of noticing the attributes of figures, grasping the intentions based on the distributions of the figures for each attribute value, and considering the meaning implied by the attributes. We developed a system for experiencing reading intentions by following our proposed method. This system also provides support functions for recognizing attributes.

Currently our research only focuses on comprehending the “intention” included in diagrams and does not support the understanding of their “contents.” The shapes or attributes of figures themselves often represent intuitive imagining of the contents. For instance, arrows may represent flow and balloons often indicate a character’s voice. To understand the diagram’s meaning, we must understand its contents. Our future work will develop a system that allows users to grasp the contents of diagrams based on the characteristics of the shapes of figures.

References

Bondy, A. S. & Frost, L. A. (1993). Mands Across the Water: A Report on the Application of the Picture-exchange Communication System in Peru: The Behavior Analyst, *16*, 123-128.

Swinkels, W., Claesen, L., Xiao, F. & Shen, H. (2018). Real-time SVM-based Emotion Recognition Algorithm: 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2), 1-6.

Chinese Grammatical Error Detection Using Adversarial ELECTRA Transformers

Lung-Hao LEE[a], Man-Chen HUNG[a], Chao-Yi CHEN[a],
Rou-An CHEN[a] & Yuen-Hsien TSENG[b*]
[a]*Department of Electrical Engineering, National Central University, Taiwan*
[b]*Graduate Institute of Library and Information Studies, National Taiwan Normal University, Taiwan*
*samtseng@ntnu.edu.tw

Abstract: We explore transformer-based neural networks for Chinese grammatical error detection. The TOCFL learner corpus is used to measure the model capability of indicating whether a sentence contains errors or not. Experimental results show that ELECTRA transformers which take into account both transformer architecture and adversarial learning technique can achieve promising effectiveness with an improvement of F1-score.

Keywords: Grammatical error diagnosis, adversarial learning, transformers, neural networks

1. Introduction

Chinese learners may make various kinds of grammatical errors, such as missing words, redundant words, incorrect word selection, or word ordering error, during their language acquisition process. An automated system able to detect such errors would facilitate Chinese learning. Previous Chinese grammatical error detection approaches were based on linguistic rules (Lee et al., 2013), machine learning classifiers (Liu et al., 2016), or their hybrid methods (Lee et al., 2014). Deep learning approaches had also been applied to detect Chinese grammatical errors (Lee et al, 2017; 2020). Recently, novel transformer-based network architectures (e.g., BERT, RoBERTa, and XLNet) achieve dominating results in many natural language processing tasks. This trend motivates us to explore transformer-based neural networks to detect Chinese grammatical errors.

This study describes our application of ELECTRA transformer architecture (Clark et al., 2020) for Chinese grammatical error detection. The TOCFL learner corpus (Lee et al., 2018) is used to evaluate the performance. Compared with previous approaches on the same dataset, ELECTRA transformers achieved an impressive improvement of F1-score which considers both detection precision and recall at the same time.

2. ELECTRA Transformers

Figure 1 illustrates our adapted ELECTRA model (Clark et al., 2020) for Chinese grammatical error detection. ELECTRA (Efficiently Learning as Encoder that Classifiers Token Replacements Accurately) is a new pre-training approach that aims to match or exceed the downstream performance of a Masked Language Modeling (MLM) pre-trained model while using less computational loading (Clark et al., 2020). During the training phase, ELECTRA trains two transformer models: the generator, which replaces the tokens in a sequence for training a masked language model; and the discriminator, which tries to identify which tokens in the sequence were replaced by the generator. If a sentence contains at least one grammatical error judged by a human, its class is labeled as 1, and 0 otherwise. The Word2Vec embedding (Mikolov et al., 2013) is used to represent sentences. All the sentences with their labeled classes are used to train our adapted ELECTRA model to automatically learn all the corresponding parameters. To classify a sentence during the testing phase, the sentence unseen in the training phase goes through the ELECTRA architecture to yield a probability value corresponding to each class. The class with the larger probability will be returned as the prediction result.

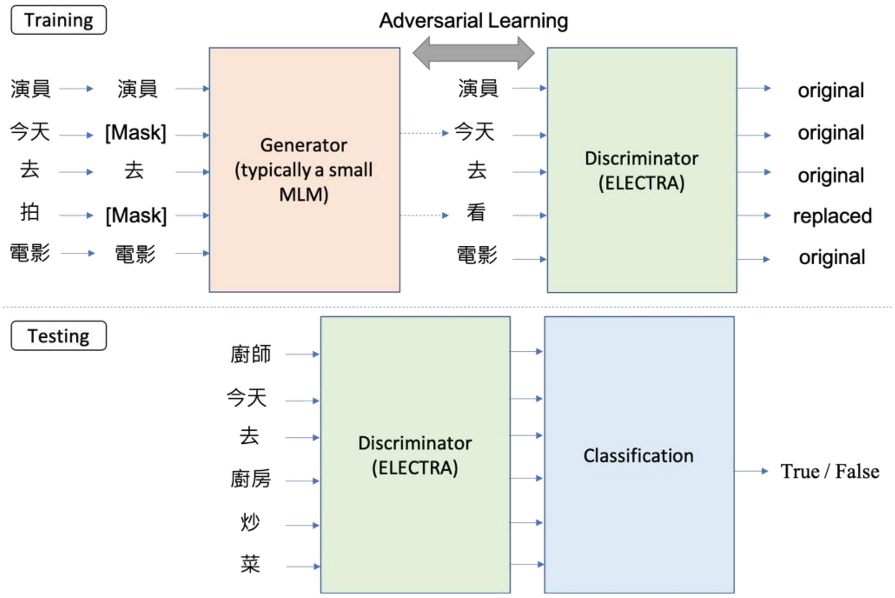


Figure 1. The illustration of ELECTRA Transformers for Chinese Grammatical Error Detection.

3. Experiments and Results

The experimental data came from the TOCFL learner corpus (Lee et al., 2018), including grammatical error annotation of 2,837 essays written by Chinese language learners originating from 46 different mother-tongue languages. Each sentence in each essay is manually labeled. This yields an annotated corpus having a total of 25,057 sentences containing at least one grammatical error, while the remaining 63,446 sentences being grammatically correct (an unbalanced distribution with 28.31% sentences having grammatical errors).

The following detection methods were compared to show their performance. (1) CNN-LSTM (Lee et al., 2017): this method integrated Convolution Neural Networks (CNN) with Long Short-Term Memory (LSTM). (2) MC-CNN-BiLSTM (Lee et al., 2020): this model had two main parts, including the multi-channel embedding representation and a CNN along with a Bidirectional LSTM (BiLSTM). (3) BERT (Devlin et al., 2018): this model was a bidirectional transformer encoder pretrained using combination of masked language modeling objective and next sentence prediction. (4). RoBERTa (Liu et al., 2019): it's a replication study of BERT pre-training that modifies key hyperparameters, removes the next-sentence pre-training objective, and applies training with much larger mini-batches and learning rates. (5) XLNet (Yang et al., 2019): this is an extension of the Transformer-XL model pretrained using an autoregressive method to learn bidirectional contexts by maximizing the expected likelihood over all permutation of the input sequence factorization order. (6). ELECTRA (Clark et al., 2020): this is our used model based on adversarial learning for Chinese grammatical error detection.

Five-fold cross validation evaluation was adapted. For Word2vec embedding representation, the whole Chinese Wikipedia (zh_tw version on Dec. 24th, 2019) was firstly segmented into words and then the segmented sentences were used to train 300 dimensional vectors for 849,217 distinct words. The hyperparameters of CNN-LSTM and MC-CNN-BiLSTM were set up according to their suggestions (Lee et al., 2017; 2020). Pretrained Chinese transformer-based models (i.e. BERT-wwm, RoBERTa-wwm, XLNet-mid and ELECTRA-base) were downloaded from HuggingFace. The configured hyper-parameters were as follows: training batch size 128, learning rate 4e-5, and max sequence length 128. F1-score, which is a harmonic mean of the precision and recall, was used as the main evaluation metric to measure the performance.

Table 1 shows the results. The four transformer-based models achieved better F1-scores that outperformed two stack-based methods. Comparing ELECTRA with the other three transformers, the former achieved F1-score of 0.6353 which is an improvement over the latter (F1-score around 0.56). It reveals that adversarial learning can enhance the performance. Besides, it's noted that ELECTRA has similar precision and recall performance without a clear bias.

]

Table 1. *Evaluation on Chinese Grammatical Error Detection*

Method		Precision	Recall	F1
Stack-based	CNN-LSTM	0.3812	0.6544	0.4808
	MC-CNN-BiLSTM	0.3669	0.7845	0.4987
Transformer-based	BERT	0.6593	0.4725	0.5501
	RoBERTa	0.6691	0.4830	0.5606
	XLNet	0.6436	0.4928	0.5580
	ELECTRA	0.6406	0.6303	0.6353

4. Conclusions

This study explores the transformed-based neural networks for Chinese grammatical error detection. We use the TOCFL learner corpus to demonstrate the model performance. The ELECTRA model, which is a transformer network architecture along with adversarial learning technique, achieved an improvement of F1-score 0.6353 for detecting whether a given Chinese sentence contains any grammatical errors or not, which is the best performance in this task as far as we know.

Acknowledgements

This study was partially supported by the Ministry of Science and Technology, under the grant 106-2221-E-003-030-MY2, 108-2218-E-008-017-MY3 and 109-2410-H-003-123-MY3.

References

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: pre-training text encoders as discriminators rather than generators. *Proceedings of ICLR'20*, https://arxiv.org/abs/2003.10555

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint, https://arxiv.org/abs/1810.04805

Lee, L.-H., Chang, L.-P., Lee, K.-C., Tseng, Y.-H. & Chen, H.-H. (2013). Linguistic rules based Chinese error detection for second language learning. *Proceedings of ICCE'13* (pp. 27-29), Bail, Indonesia: Asia-Pacific Society for Computers in Education.

Lee, L.-H., Lin, B.-L., Yu, L.-C., & Tseng, Y.-H. (2017). Chinese grammatical error detection using a CNN-LSTM model. *Proceedings of ICCE'17* (pp. 919-921), Christchurch, New Zealand: Asia-Pacific Society for Computers in Education.

Lee, L.-H., Tseng, Y.-H., & Chang, L.-P. (2018). Building a TOCFL learner corpus for Chinese grammatical error diagnosis. *Proceedings of LREC'18* (pp. 2298-2304), Miyazaki, Japan: ACL Anthology.

Lee, L.-H., Wang, Y.-S., Lin, P.-C., Hung, C.-T., & Tseng, Y.-H. (2020). Multi-channel CNN-BiLSTM for Chinese grammatical error detection. *Proceedings of ICCE'20* (pp. 558-560), online: Asia-Pacific Society for Computers in Education.

Lee, L.-H., Yu, L.-C., Lee, K.-C., Tseng, Y.-H., Chang, L.-P., & Chen, H.-H. (2014). A sentence judgment system for grammatical error detection. *Proceedings of COLING'14* (pp. 67-70), Dublin, Ireland: ACL Anthology.

Liu, Y., Han, Y., Zhou, L., & Zan, H. (2016). Automatic grammatical error detection for Chinese based on conditional random field. *Proceedings of NLPTEA'16* (pp. 57-62), Osaka, Japan: ACL Anthology.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint, https://arxiv.org/abs/1907.11692

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of NIPS'13* (pp. 1-10), Stateline, Nevada.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. & Le, Q. V. (2019). XLNet: generalized autoregressive pretraining for language understanding. arXiv preprint, https://arxiv.org/abs/1906.08237

Presentation Scenario Design Support System That Prompts Awareness of Other Viewpoints

Kazumi MASAKADO*, Yuki HAYASHI & Kazuhisa SETA
Graduate School of Humanities and Sustainable System Sciences, Osaka Prefecture University, Japan
*masakado@ksm.kis.osakafu-u.ac.jp

Abstract: Even when a learner (novice researcher) has thoroughly reviewed their presentation materials, it is not uncommon for others (experts) to find that there is still room for improvement. This is commonly because learners do not sufficiently engage in metacognitive activities such as reviewing their thoughts or shaping their knowledge from other viewpoints. Therefore, in this study, we focus on using the task of creating presentation materials as an opportunity to promote awareness of other viewpoints, and we propose an information system mechanism that gives advice on how to correct and improve presentation materials based on audience models and presentation scenarios.

Keywords: Presentation scenario, cognitive conflict, awareness of other viewpoint

1. Introduction

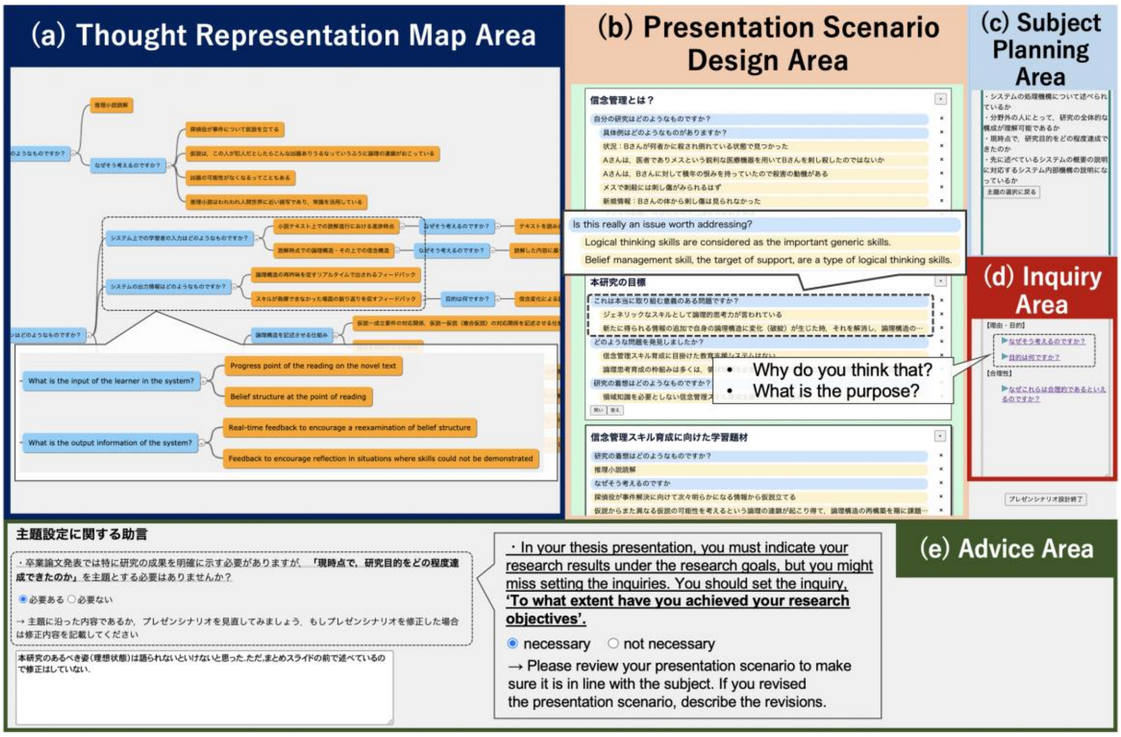
In this study, we examine how learning support systems can help prepare presentation scenarios that facilitate meaningful knowledge creation opportunities, and we examine the required functions to achieve this. For example, when making a presentation for a research topic, it is not always easy for learners (novice researchers) to state their research topic academically and prepare presentation materials that clearly express the value and content of the topic to the audience. In such situations, the learner gets help from research collaborators (e.g., seniors, peers, professors, etc.) to resolve possible issues and create sophisticated presentation materials in a trial-and-error way.

So far, several studies have proposed the learning support system that promotes the quality of learners' presentation scenarios based on general presentation structures. For example, Tanida, Hasegawa, & Kashihara (2008) proposed a method to encourage planning and reflecting on the presentation document by explicitly assigning the presentation structure to the learner's own slides. The presentation structure is systematized based on typical presentation structures used in presentation materials created in the laboratory, such as 'outline,' 'research objectives,' and 'approach.' Kojiri, & Watanabe (2016) proposed a learning support system to organize the presentation contents by annotating a specific topic to each content. The topic model used in the system consists of the logical relationships among the typical presentation components, such as 'problem,' 'purpose,' and 'method.' While these approaches help learners to understand the logical structure of their presentation contents rather general, research context independent level, they do not give advice from audience's viewpoints capturing learners' research content to help examination of the presentation scenario.

Generally, presentation design activities can be difficult due to the complex tasks required and the fact that the processes for handling them cannot be specified (i.e., ill-defined problem) (Jonassen, 1997). Therefore, the goal of this study is to examine how to transition presentation preparation efforts (including discussions) from confusing and ad-hoc, as pointed out by Hollnagel (1996), to strategic and knowledge-creating.

2. Presentation Scenario Design Support System

We developed a support system for presentation scenario design that cooperates with the "Forest" internal conversation support system developed by Mori et al. (2019). Figure 1 shows our system


(a) Thought Representation Map Area
推理小説読解
なぜそう考えるのですか？
探偵役が事件について仮説を立てる
仮説は、この人が犯人だとしたらこんな凶器ありうるなっていうふうに論理の連鎖がおこっている
凶器の可能性がなくなるってこともある
推理小説はわれわれ人間世界に近い描写であり、常識を活用している
システム上での学習者の入力はどのようなものですか？
小説テキスト上での読解進行における進捗時点
読解時点での論理構造・その上での信念構造
システムの出力情報はどのようなものですか？
論理構造の再吟味を促すリアルタイムで出されるフィードバック
スキルが発揮できなかった場面の振り返りを促すフィードバック
目的は何ですか？
論理構造を記述させる仕組み
What is the input of the learner in the system?
Progress point of the reading on the novel text
Belief structure at the point of reading
What is the output information of the system?
Real-time feedback to encourage a reexamination of belief structure
Feedback to encourage reflection in situations where skills could not be demonstrated
(b) Presentation Scenario Design Area
(c) Subject Planning Area
(d) Inquiry Area
(e) Advice Area
問い
答え
主題の選択に戻る
プレゼンシナリオ設計終了

信念管理とは？
自分の研究はどのようなものですか？
具体例はどのようなものがありますか？
状況：Bさんが何者かに殺され倒れている状態で見つかった
Aさんは、医者でありメスという鋭利な医療機器を用いてBさんを刺し殺したのではないか
Aさんは、Bさんに対して積年の恨みを持っていたので殺害の動機がある
メスで刺殺には刺し傷がみられるはず
新規情報：Bさんの体から刺し傷は見られなかった

Is this really an issue worth addressing?
Logical thinking skills are considered as the important generic skills.
Belief management skill, the target of support, are a type of logical thinking skills.

本研究の目標
これは本当に取り組む意義のある問題ですか？
ジェネリックなスキルとして論理的思考力が言われている
新たに得られる情報の追加で自身の論理構造に変化（破綻）が生じた時，それを解消し，論理構造の…
どのような問題を発見しましたか？
信念管理スキル育成に目掛けた教育支援システムはない
論理思考育成の枠組みは多くは，領
研究の着想はどのようなものですか？
領域知識を必要としない信念管理ス

- Why do you think that?
- What is the purpose?

信念管理スキル育成に向けた学習題材
研究の着想はどのようなものですか？
推理小説読解
なぜそう考えるのですか
探偵役が事件解決に向けて次々明らかになる情報から仮説立てる
仮説からまた異なる仮説の可能性を考えるという論理の連鎖が起こり得て，論理構造の再構築を陽に課題…

・システムの処理機構について述べられているか
・分野外の人にとって，研究の全体的な構成が理解可能であるか
・現時点で，研究目的をどの程度達成できたのか
・先に述べているシステムの概要の説明に対応するシステム内部機構の説明になっているか

【理由・目的】
▶なぜそう考えるのですか？
▶目的は何ですか？
【合理性】
▶なぜこれらは合理的であるといえるのですか？

主題設定に関する助言

・卒業論文発表では特に研究の成果を明確に示す必要がありますが，**「現時点で，研究目的をどの程度達成できたのか」**を主題とする必要はありませんか？

● 必要ある ○ 必要ない

→ 主題に沿った内容であるか，プレゼンシナリオを見直してみましょう．もしプレゼンシナリオを修正した場合は修正内容を記載してください

本研究のあるべき姿（理想状態）は語られないといけないと思った.ただ,まとめスライドの前で述べているので修正はしていない.

・In your thesis presentation, you must indicate your research results under the research goals, but you might miss setting the inquiries. You should set the inquiry, **'To what extent have you achieved your research objectives'.**

● necessary ○ not necessary

→ Please review your presentation scenario to make sure it is in line with the subject. If you revised the presentation scenario, describe the revisions.

Figure 1. Presentation Scenario Design Support System.

implemented as a web application. The system consists of a (a) thought representation map area, a (b) presentation scenario design area, a (c) subject planning area, an (d) inquiry area, and an (e) advice area.

In their daily research activities, learners organize their research content as a (a) thought representation map using the Forest. In the Forest, learners externalize their thoughts as a chain of questions and answers (pyramidal structure (Minto, 2009)) and organize their thoughts by engaging in internal conversation. Learners can select and place blue inquiry nodes from the (d) inquiry area.

In the presentation design activity, learners should consider the audience and dig deeper into issues that should be focused on according to the subject. To promote this activity, learners can show the (b) presentation scenario design area and (c) subject planning area, when designing a presentation. In the subject planning area, the audience's viewpoints (e.g., "Is the academic significance mentioned" and "Is the technical background mentioned") are presented so learners can consciously select the subjects to be mentioned in the presentation. Learners specify the corresponding content of each frame for each slide as a question and answer structure to be conveyed to the audience. If new inquiries (i.e., cognitive conflicts) arise during this process, learners can reflect them in the thought representation map and reconstruct them to examine their answers, thereby resolving cognitive conflicts.

When learners judge they could have designed a well-thought-out presentation, the system generates advice based on other viewpoints in response to requests from the learner. The learners' judgments of the decision-making about the advice (i.e., whether the learner reflects the advice or not with the reason) and the content reflected in the presentation scenario are recorded in a log file that can be used for discussion with their supervisors.

3. Case Study: Practical Use in Presentation Scenario Design Activity

Our system has been used in authentic activity to make presentation scenarios for bachelor and master theses to confirm whether our system promotes awareness of other viewpoints, helps reconstruct thoughts, and facilitates knowledge-creating discussions with supervisors. Five students in our laboratory (four undergraduates and one master student) have used the system in practice.

In the scenario design process, the learner added many new inquiries and answers in the thought expression map, indicating that he delved into the subjects focused on in the scenario. Interactions with the system (i.e., advice given by the system and the learner's explicit judgments on it) were as follows:

(1) **Advice given about reexamining the subjects of the presentation scenario:** The system advised, "*In your master thesis presentation, you must indicate your research results under the research*

goals and set the inquiries. You should set the inquiry, 'To what extent have you achieved your research objectives set for this presentation.'"

Learner's judgment: The learner selected "*necessary*" and stated, "*I thought I should show the concrete image of our research when we achieve our research goal. Nevertheless, I did not modify it because I mentioned it in the slide entitled 'Concluding remarks.'*"

(2) **Advice given about reexamining the presentation scenario:** The system advised, "*The inquiry, 'What are your future tasks?' should be included and answered by considering the inquiry, 'To what extent have you achieved your research objectives set for this presentation' You should review this based on your research objectives.*"

Learner's judgment: The learner selected "*Do not review*" and stated the reason as "*Already described.*"

(3) **Advice given about reflecting on the learner's judgment process to increase readiness for discussion:** The system asked, "*What did you learn at this time?*" and showed the following decision-making process: *"You judged it necessary to respond to, 'In your master thesis presentation, you must indicate your research results under the research goals and set the inquiries. You should set the inquiry, 'To what extent have you achieved your research objectives set for this presentation' Then, you revised Frame No. 12 entitled 'Concluding remarks,' adding the inquiry, 'To what extent have you achieved the research objectives set for this presentation' and mentioning, 'I added an explanation of the experiment results consistent with the purpose of the research.'"*

Learner's answer: The learner stated, "*I realized that I had not been aware that the current status and my research results were not clearly stated, which is required in a master thesis presentation.*"

After finishing learners' scenario design process, they discussed their designed presentation scenarios with their supervisors by referring to their decision-making log described above. According to the questionnaire survey conducted after the discussion, learners found the system quite helpful in preparing for their discussions. Some example responses were: "The system pointed out what was missing or what I had forgotten should be embodied in my presentation scenario, so I was able to add the topic to the agenda for discussion." Although there have been only five case studies, we feel the system encourages the recognition of other's viewpoints and helps prepare meaningful opportunities for collaborative knowledge-building.

4. Conclusion

We developed a system that promotes knowledge-building in research presentations through learner tasks such as organizing research content in a thought expression map and designing a presentation scenario using the thought representation map. The practical study suggests that the presentation scenario design using the system encouraged creative knowledge-building discussions.

References

Hollnagel, E. (1993). Human reliability analysis: context and control. Academic Press.

Jonassen, D. H. (1997). Instructional design models for well-structured and ill-structured problem-solving outcomes. *Educational Technology Research and Development, 45*(1), 65–94.

Kojiri, T. & Watanabe, Y. (2016). Contents Organization Support for Logical Presentation Flow. *Workshop Proceedings of the 14th Pacific Rim International Conference on Artificial Intelligence*, 145–156.

Minto, B. (2009). The Pyramid Principle: Logic in Writing and Thinking. Pearson Education.

Mori, N., Hayashi, Y., & Seta, K. (2019). Ontology Based Thought Organization Support System to Prompt Readiness of Intention Sharing and Its Long-term Practice. *The Journal of Information and Systems in Education, 18*(1), 27–39.

Tanida, A., Hasegawa, S., & Kashihara, A. (2008). Web 2.0 services for presentation planning and presentation reflection. *Proc. International Conference on Computers in Education*, 565–572.

Visualization of Topics and Logical Development Based on Reader's Understanding for Reading Support

Yuki OKANIWA[a*] & Tomoko KOJIRI[b]
[a]*Graduate School of Science and Engineering, Kansai University, Japan,*
[b]*Faculty of Engineering Science, Kansai University, Japan*
*k595436@kansai-u.ac.jp

Abstract: When reading editorials, understanding the meaning of sentences and their relations is essential for grasping topics and their logical structures. If inter-sentence relations are not understood correctly, the topics and logical structures derived from them will not follow the structure of the editorials. That is, the derived topics do not correspond to those of the editorials' paragraphs. The derived logical development does not conclude the main opinion. This study supports the accurate understanding of inter-sentence relations by notifying readers of inaccurate topics and the logical development derived by their understanding. Our developed system visualizes the correspondence between paragraphs and derived topics as well as the reverse flow of the logic from the opinions to investigate whether all sentences are constituents of the opinions.

Keywords: Reading support system, logical structure, visualization, inter-sentence relations

1. Introduction

When reading editorials, readers must first understand the meaning of each sentence and the relations among all sentences. Based on the relations, they grasp the topics, the logical structure, and the opinion. Understanding inter-sentence relations is essential for grasping opinions. This paper supports readers who are struggling to read inter-sentence relations correctly to understand topics and their logical structure or to grasp the ideas.

As support for improving reading understanding skills, Fukunaga et al. developed a system in which readers underlined the parts of a text that express important topics and receive feedback based on a comparison of the underlined parts and the correct answers (Fukunaga et al., 2005). This system demonstrates the accuracy of the understood topics without providing a method of accurate understanding.

Fukumoto et al. argued that inter-sentence relations must be correctly understood to grasp the opinions of editorial texts (Fukumoto & Tsujii, 1994). For improving the reading understand of inter-sentence relations, Mochizuki et al. provided an environment in which the relations between sentences are organized to promote structural understanding (Tsubakimoto et al., 2008). They exploited the organization of inter-sentence relations to derive opinions without supporting the understanding of correct inter-sentence relations.

Our study promotes an accurate understanding of inter-sentence relations. We develop a system that visualizes logical structure and distributes topics in paragraphs that can be inferred from the inputted inter-sentence relations understood by readers. By monitoring the visualization, readers are expected to revise their interpretations of inter-sentence relations based on the visualized information and derive an accurate understanding.

2. Reading Understand Support System

Readers start at the level of sentences and grasp the writer's opinion in the following process:

1. Understand the sentences and their inter-sentence relations;
2. Grasp the topics;
3. Compose a logical structure from the relations among the topics;
4. Derive the opinion from its logical structure.

In step 1, readers understand the meaning of sentences based on words, grammar, and the inter-sentence relations based on the conjunctions. In step 2, they grasp topics based on the inter-sentence relations. They also grasp the logical structure of topics based on the inter-sentence relations in step 3. Logic is developed in such a way to lead them to derive a writer's opinion in step 4. Among these steps, step 1 is essential, since steps 2 to 4 are based on the results of step 1. If readers are unable to correctly read the inter-sentence relations, they will not correctly understand the topics and logical structure. Hence, the opinion will not be understood.

If topics are not grasped accurately, paragraphs may not correspond to the topics. If the logical relations are not grasped sufficiently, some topics may not lead to the opinions derived by readers. If readers notice such conflicts, they may read the sentences carefully and find appropriate inter-sentence relations. This study proposes a reading support system of inter-sentence relations that gives awareness of the inaccuracy of the understood inter-sentence relations. Our system visualizes the topics and logical structures that can be recognized based on the inter-sentence relations understood by readers and encourages them to recognize the inaccuracy by themselves.

Figure 1 overviews the system, which has an editorial database that stores such texts to show to readers. Readers select a file name from the database and start their reading practice. The system has two interfaces. One is an inter-sentence relation input interface where readers can input the inter-sentence relations that they read from the editorials. This system provides three types of relations: causal, generalization/ specification, and supplemental. The inputted inter-sentence relations are stored in the learning-log database. The other interface is visualization. It grasps the topics and logical structures that can be inferred from the inputted inter-sentence relations and presented to readers based on requests.

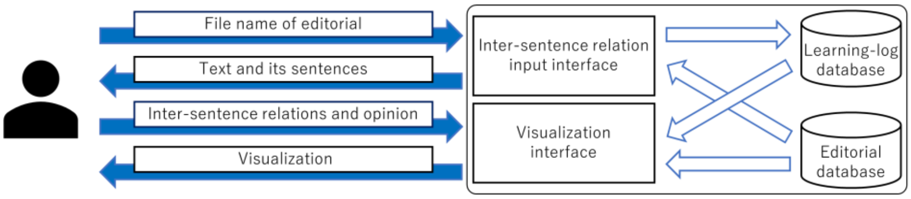


Figure 1. System overview.

3. Visualization Interface

The visualization interface shows the topics, the logical structure, and logical development that can be grasped by the inter-sentence relations understood by the reader. It provides two visualization forms: the logical development visualization and the topic division visualization.

The logical development visualization shows the logical structure that can be grasped by readers and the logic toward an opinion. The logical structure is the relations among topics. This study defines a logical structure map to express the logical structure. Figure 2 shows the configuration of a logical structure map. Nodes correspond to sub-topics, and links show the relations among topics. The vertical axis represents the level of the details of the sub-topics, and the horizontal axis represents the logical development. Since the sentences of the supplemental relations indicate identical sub-topics, these sentences are gathered to form one node. Sentences without supplemental relations indicate that they themselves represent sub-topics, so they form nodes on their own. Links show either causal relations or generalization/specification relations between two sub-topics. For sub-topics that are connected by causal relations, the result topics are arranged as right nodes along the horizontal axis, and the links are directed from the cause to result nodes. Specialized sub-topics may belong to topics that resemble abstract ones, so they are arranged as the lower node along the vertical axis and generalized topics as the upper node, and the links are directed from the lower to upper nodes.

Figure 3 shows another example of a logical structure map. Since the causal relation between topics 1 and 3 is missing, there are two chunks. Such structure indicates that, according to the understanding of readers, more than one logical structure exists in the target editorial.

Logical development should be designed to derive an opinion to which all topics should have a path. Therefore, it is useful to check whether all topics are included in all the paths from an opinion. Our interface creates pictures showing like an animation for following the link in a reverse direction from the opinion node by adding red to each node. Figure 4 shows some of pictures when topic 7 is read as an opinion. First, topic 7 is colored, and then its detailed topics are colored (Figure 4a). Next, the cause of topic 7, such as topic and its detailed topics, are colored (Figure 4b). In the same way, topics 1 and 2 are colored (Figure 4c). Since all the topics are successfully colored in Figure 5, this logical structure is probably appropriate. If some nodes become uncolored, the inter-sentence relations correspond to the uncolored topics may be grasped inaccurately.

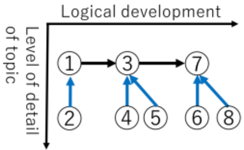


Figure 2. Example of Logical Structure Map.

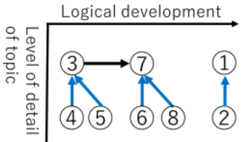


Figure 3. Example of Improper Logical Structure

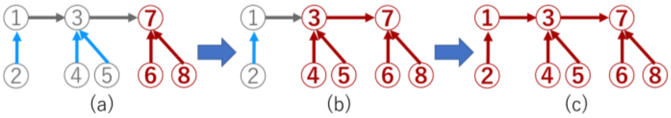


Figure 4. Example of Coloring Nodes from Opinion along Reverse Direction of Logical.

On the other hand, the topic division visualization shows the distribution of the topics for each paragraph. Figure 5a shows how the sentences are arranged by paragraphs whose colors are assigned for individual topics. Since specialized nodes may represent the same topics with their upper nodes, sentences that belong to the same sub-trees are assigned the same color. If the same color is distributed over several paragraphs, or if a paragraph is changed within a topic, as shown in Figure 5b, it can be suggestive that the reader's understanding of the topic may be inaccurate.

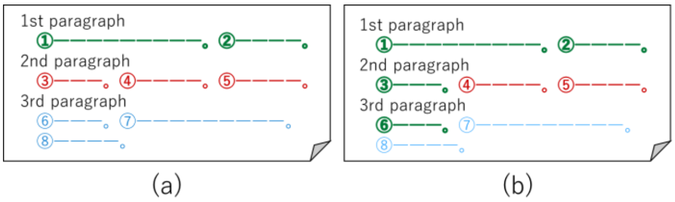


Figure 5. Imagination of Topic Division Visualization.

4. Conclusion

We developed a system that supports readers to understand the inter-sentence relations for the purpose of accurate understanding the opinions of the editorials. Our system visualizes the logical developments and the distribution of topics based on reader understandings of inter-sentence relations to make them aware of their misunderstandings and to revise their understanding of the opinions and its reasons.

In this study, we support the detection of mistakes in inter-sentence relations, but we do not support the correction of the understanding of inter-sentence relations. As a future work, it is necessary to propose a method to support correction by feedback on mistakes in inter-sentence relations.

References

Fukunaga, Y., Hirashima, T., & Takeuchi, A. (2005). Implementation and Effectiveness of a Feedback Feature in Underlining to Promote Reading Comprehension of e-Learning Instructional Materials. *In Proceedings of International Conference on Computers in Education*, pp. 101-108.

Fukumoto, F., & Tsujii, J. (1994). Automatic Recognition of Verbal Polysemy. *In Proceedings of The 15th International Conference on Computational Linguistics,* Volume 2, pp. 762-768.

Tsubakimoto, M., Mochizuki, T., Nishimori, T. et al. (2008). The Impact of Making a Concept Map for Constructive Reading with the Critical Reading Support Software "eJournalPlus." *In E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education,* pp. 506-514.

Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Flip & Slack - Active Flipped Classroom Learning with Collaborative Slack Interactions

Kyong Jin SHIM*, Swapna GOTTIPATI & Yi Meng LAU
School of Computing and Information Systems, Singapore Management University, Singapore
*kjshim@smu.edu.sg

Abstract: Active flipped classroom learning is stipulated with faculty structuring the activities involving constructive interactions, either formal or informal. Sharing ideas and responding to ideas improve the cognitive skills of the students. Encouraging peers to contribute to class activities and respecting peers contribute to the development of affective skills. We present an integrated platform for cognitive and affective skills development. A flipped classroom arrangement allows the faculty to focus more on in-class activities such as programming and lab exercises to support active learning in computing courses. We share the design of an innovative flipped classroom model integrated with Slack and present the design principles applied to emphasize cognitive and affective skills development through flip-slack model. We evaluate the model on the second-year computing course, Web Application Development II. We share the findings of this case study by analyzing student feedback and their grades.

Keywords: Flipped classroom, active learning, collaborative learning, slack

1. Introduction

Active learning research shows several advantages, and the spectrum includes three popular approaches: collaborative learning, cooperative learning, and problem-based learning (Davidson & Major, 2014; Caceffo et al., 2018; Fassbinder et al., 2015). An organized and strong collaboration between students and instructor is required to achieve the full potential of this approach. Moreover, the instructor's role shifts from knowledge disseminator to learning facilitator. At the same time, students ought to accept more responsibility for their own learning and adapt to a classroom style that is different from that experienced in previous courses. During the learning process, the students develop both cognitive and affective skills which are important for the digital era (Yamamoto & Ananou, 2015; Gokarn et al., 2019). Flipped classroom model enables active learning with the components supporting all the three phases of the spectrum (Maher et al., 2015; Thongmak, 2017). Active learning is stipulated with faculty structuring the activities involving constructive interactions, either formal or informal. Interactions such as sharing ideas and responding to ideas improve the cognitive skills of the students while interactions such as encourage peers on contributions and respecting peers contribute to the development of affective skills (Thongmak, 2017).

Teaching computing courses comes with its own unique set of challenges. The nature of most computing courses requires deep hands-on "practices" of concepts – and this has been the norm across many science and engineering courses around the world (Brown & Atkins, 1996). For learners to truly master the concepts and principles of computing, it is critical that they can put the concepts and principles into solution development. Along with the cognitive skills, the computing students are required to develop the affective skills as there a strong correlation to their learning performance (Wu et al., 2016). Thus, in our school, computing courses are designed to mix both concepts and hands-on lab exercises in a span of three hours of lesson time with a small classroom setting of 45 students to enable strong in-class interactions.

Flipped classrooms are adopted in various courses and at various levels of higher education (Maher et al., 2015; Gehringer & Peddycord, 2013; Lockwood & Esselstein, 2013; Shim et al., 2021). Computing courses have benefited more from this approach due to nature of the content which is information-intensive and skill based. The model enables to develop the cognitive and affective skills efficiently as well (Foldnes, 2016; Mentz et al., 2008). The learning is not only in term of understanding and remembering but also applying and creating solutions to the problems. The courses are designed with lectures, in-class activities, and projects similar to the real-world cases (Fassbinder et al., 2015).

A flipped classroom arrangement allows the computing faculty to focus more on the in-class activities such as programming and lab exercises to support active learning (Caceffo et al., 2018). It enables the collaborative student learning environment through the group activities such as class diagrams or database modelling, etc. (Lockwood & Esselstein, 2013). Cooperative learning which is a structured process in which team members work towards accomplishing a common goal, stressing positive interdependence, and individual and group accountability is crucial for computing courses. It is achieved with in-class activities such as pair programming or design the web page (Foldnes, 2016; Mentz et al., 2008). Finally, in flipped class, some faculty also integrate Problem-Based-Learning where students are organized into small groups with the goal of solving problems that have some similarity with real world problems such as registration web page creation, creating library database etc., and faculty facilitate the process (Fassbinder et al., 2015).

Interactions in flipped classroom are enabled manually or with technologies such as discussion forums from learning management system or social media such as wiki pages (Hardaway & Scamell, 2005). Computing courses requires a continuous support from the faculty and peer to help fix the bugs in the code or share the algorithms to solve the problems. The support should be provided not only in the classroom but also for the out-of-class activities (Bergmann & Sams, 2012). This requires an efficient and simple interaction platform for faculty and students to discuss the topics and the knowledge is managed with features like teams, search, file sharing, annotations, and audio-video chats. LMS platforms are commonly used for interactions but are limited on the innovative features (Al-Ibrahim, & Al-Khalifa, 2014). Social media provides such free platforms with innovative features such as wiki, Google hangouts, Slack and MS teams are the popular ones to name (Ferreira, 2014; Tuhkala & Kärkkäinen, 2018). Moreover, the Gen-Z students are more comfortable using social media platforms.

In this paper, we share the design of innovative model, flipped classroom integrated with Slack. We focus on the design principles that we applied to emphasize the cognitive and affective skills development thorough the flip-slack model. We evaluate the model on the second-year computing course, Web Application Development II. The impact of the models on students' learning process and the development of cognitive and affective skills was examined. A questionnaire based on the learning theories was developed to evaluate the effectiveness of the model. We share the evaluations of our model by analyzing the student survey, student grades, and lessons learnt.

2. Background

2.1 Cognitive Skills

In flipped classroom, the cognitive skills development methods are also flipped (Eppard, & Rochdi, 2017). Remembering, understanding, and applying are usually achieved out of classroom. The remaining skills are achieved in-class. Instructional methods for in-class are problem-based activities, discussions, and facilitation. We identified typical activities in our school's computing courses and mapped them to different cognitive levels (Table 1).

Table 1. *Cognitive Levels and Activities for Computing Courses*

Cognitive Levels	Achievable Time	Example Activities
Remembering, understanding, & applying	Pre-class	Watching videos, reading slides, attempting quizzes etc.
Analyzing, evaluating, & creating	In-class	Doing quizzes, problems, advanced concepts, & collaborative learning
Evaluating & creating	Post-Class	Extra Exercises, doing course project

2.2 Affective Skills

In flipped classroom, affective skills are mainly developed through interactions. Affective outcomes refer to educational outcomes regarding students' satisfaction, confidence, motivation, emotions, attitudes, and feelings toward learning, the subject matter itself, or educational activities (Krathwohl et al., 1964; Anderson et al., 2001). Affective domain includes five levels. We identified typical activities in our school's computing courses and mapped them to different affective levels (Table 2).

Table 2. *Affective Levels and Activities for Computing Courses*

Level	Description	Example Activities
Receiving	Awareness of the need and willingness to hear selected attention	Watching videos, listening to instructor attentively
Responding	Actively participate in learning and responding to various appearances	Participation in class discussions, presentations, asking questions
Valuing	Ability to judge the worth or value of objects or phenomena and express it	Proposing plans to improve team skills, supporting ideas from the team
Organization	Ability in putting together the ideas, information, and create a value system	Comparing the ideas, relate to the main problem, prioritize time effectively
Characterization	At the highest level the value has been "internalized"	Work efficiently independently or in teams, and practices ethics.

2.3 Flipped Classroom Basic Model

Theoretical frameworks help the faculty to design the flipped classroom model suitable for their course and are broadly categorized into two types: student-centered learning theories and teacher-centered learning theories. Most popular theories applied for such designs include, Bloom's taxonomy, active learning, problem-based learning, collaborative learning, experiential learning, theory of reasoned action, self-determination theory, behaviorist theory, and constructivist theory. These theories aim at the cognitive and affective skills development in a student to achieve the better outcomes of the flipped classroom (Bishop et al. 2013). Table 3 shows the components, examples, and benefits of each component in the flipped classroom.

Table 3. *Flipped Classroom Components*

Components	Examples	Benefits for flipped class
Student-centered learning theories	- Constructivist theory (Vygotsky, 1978; Grabinger & Dunlap, 1995), - Peer-assisted learning (Topping & Ehly, 1998)	- Critical thinking - Communication and collaboration skills - Motivation and responsibility
Teacher-centered learning theories	- Behaviourist theory (Watson, 1920; Skinner, 1948)	- Expert knowledge - Organized and disciplined - Effective evaluations
Out-of-class activities	- Read textbook - Watch Videos - Take online quiz - Submit programs	- Personalized learning support - Mobile learning support - Autonomous learning Support - Cognitive and affective skills development - Improved self-learning ability
In-class activities	- Pair Programming Labs - Group Problem Solving Activities - Flexible Quiz Activities	- More interactions with instructor - Collaborative learning support - Cognitive and affective skills development - Increased material retention

According to Moore's theory, students construct their interactions with four types of elements, namely, peers, instructors, contents, and technologies either in-class or outside of the class time (Moore, 1989). Out of classroom interaction platforms most commonly employed in flipped classroom are LMS or social media platforms like wikis and blogs (Ferreira, 2014). Slack, MS Teams, and Google Hangouts are similar platforms, and we choose Slack due to some free features not available in other tools (Tuhkala & Kärkkäinen, 2018) such as efficient conversation segmenting, tracking and searching via custom channels, social interactions via 'mentioning' (similar to popular social media platforms such as Facebook, Instagram and Twitter), and built-in analytics.

2.4 Web Application Development Courses

In our school, all students must complete two web application development courses. Web Application Development course (WAD I) equips students with the knowledge and skills to develop database-driven web applications using HTML, PHP, and MySQL. Upon successful completion of WAD I, Web Application Development course (WAD II) teaches students how to develop well-styled and responsive web applications that provide rich user experiences using HTML, CSS, Bootstrap, JavaScript, and Vue.js. While WAD I focuses fully on back-end development centering around client-server architecture, WAD II builds on top of WAD I by teaching students how to develop full-stack web applications and also explores interactions with external APIs.

3. Flip & Slack Model

3.1 Research Questions

Diversity in students' learning styles and motivation challenges can be addressed by flip model and facilitate both personalized and collaborative learning (Goedhart et al., 2019). Our first set of questions study the impact of the flip-slack model on the learning process.

RQ1: What is the impact of flipped class on the learning process?
RQ2: What is the impact of slack interactions on the learning process?

Development of cognitive and affective skills can be achieved through successful interactions (Johnson & Johnson, 2002) and effective teaching principles (Bishop et al. 2013) employed in flipped classroom. Our second set of questions study the impact of the flip-slack model on the development of skills.

RQ3: Does the flip-slack model improve cognitive skills?
RQ4: Does the flip-slack model improve affective skills?

To assess the statistical significance of the model, we study the correlations between self-assessed skill development and students' performance on course assessments.

RQ5: What is the relationship between skills developed and student grades?

3.2 Flip & Slack Design

Based on the previous frameworks and theories, we propose flip-slack teaching model built on cognitive and affective principles (Figure 1).

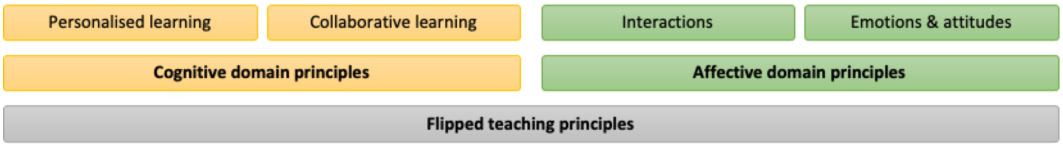

Personalised learning
Collaborative learning
Interactions
Emotions & attitudes
Cognitive domain principles
Affective domain principles
Flipped teaching principles

Figure 1. Flip-Slack Classroom Model – A Platform for Cognitive and Affective Skills Development.

Flipped classroom design aims at active (student-centered) learning, flexibility, and simplicity. The fundamental principles can be categorized under in-class and out-of-class as described below:

1. Out-of-class: Enable effective learning by using instructor created short videos that ensure Coherence, attention guiding effect, segmentation, learner control effect, engagement effect, encouraging mental model making, learner control effect, misconception effect and integrated

practice activities. Use online exercises and discussion forums to motivate students' class participation.

2. In-class: Activate student's motivation with formative assessments such as quiz and feedback on their submissions. Use group activities to enable active learning by solving varied tasks and real-world problems.

Cognitive design aims at the intellectual skills development by enforcing components of active learning in the classroom delivery as shown below:

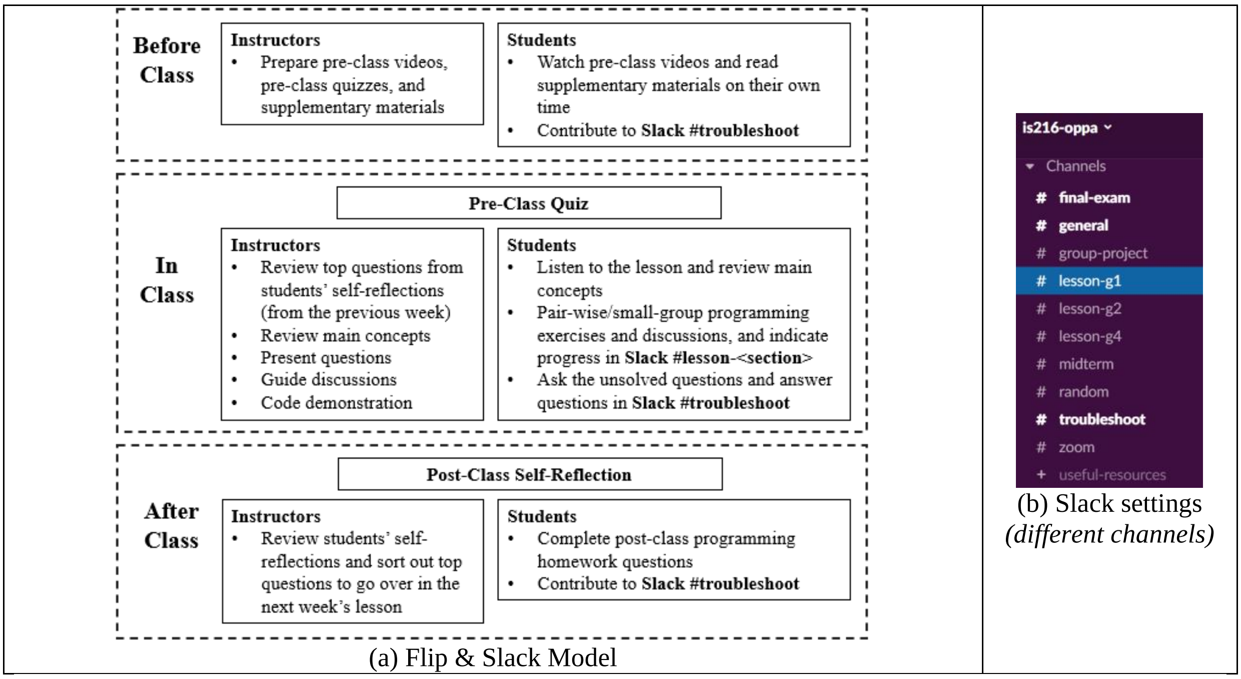
1. Before class: Focus on remembering and understanding cognitive levels by assessing the students on basic topics via online quizzes which are individual and flexible activities enabling personalized learning.
2. In class: Focus on applying and analyzing cognitive levels by organizing and assessing the students on advanced topics via group activities supporting collaborative learning.
3. After class: Focus on evaluating and creating cognitive levels by assessing the students on individual or group projects.

Affective design aims at the social skills development by enabling interactive platforms and opportunities to demonstrate appropriate emotions and attitudes.

1. Before class: Focus on receiving and responding affective levels by watching videos, reading online-discussions, and replying to the online-discussions promoting interactions.
2. In class: Focus on valuing and organization levels by participating in the group activities and discussions with the facilitators.
3. After class: Focus on characterizing level by collaborating on the large-scale course projects or assignments with strong interactions and ethical practices on the technology platforms.

3.3 Flip & Slack Model – Case Study

Figure 2 depicts the flip and slack settings for WAD II course based on the principles presented in previous section. The figure shows the flipped classroom structure and Slack channels design.



Before Class

Instructors
- Prepare pre-class videos, pre-class quizzes, and supplementary materials

Students
- Watch pre-class videos and read supplementary materials on their own time
- Contribute to **Slack #troubleshoot**

In Class

Pre-Class Quiz

Instructors
- Review top questions from students' self-reflections (from the previous week)
- Review main concepts
- Present questions
- Guide discussions
- Code demonstration

Students
- Listen to the lesson and review main concepts
- Pair-wise/small-group programming exercises and discussions, and indicate progress in **Slack #lesson-<section>**
- Ask the unsolved questions and answer questions in **Slack #troubleshoot**

After Class

Post-Class Self-Reflection

Instructors
- Review students' self-reflections and sort out top questions to go over in the next week's lesson

Students
- Complete post-class programming homework questions
- Contribute to **Slack #troubleshoot**

is216-oppa
Channels
# final-exam
# general
# group-project
# lesson-g1
# lesson-g2
# lesson-g4
# midterm
# random
# troubleshoot
# zoom
+ useful-resources

(b) Slack settings *(different channels)*

(a) Flip & Slack Model

Figure 2. Flip & Slack Settings for WAD II Course.

Each week, students are given next week's lesson videos and materials for out-of-class activities. Students study slides and watch short videos. During the lesson time, the teaching team, 1) review student feedback from the previous week, 2) review common questions from the previous week, 3) recap on the current week's concepts, 4) give in-class "hands-on" challenges to students, and 5) utilize the remaining class time for one-on-one or small group-based consultation where the teaching team members actively walk around and mentor students on coding, debugging and other course-related individual or group actives. The teaching team members alternate checking on questions posted in Slack – and provide online consultation as and when necessary.

Two types of users are created in Slack: 1) administrators (faculty and teaching assistants), 2) general users (students). Administrators can create and edit channels, manage users, configure settings (e.g. installation of extensions), and view analytics. We created a separate public channel for each topic. Channel "general" is for making announcements class announcements. Channel "troubleshoot" is for technical troubleshooting where students and teaching team members post questions and answers concerning hands-on lab exercises. During the class time, section-specific channels (e.g. lesson-g1 is for section G1) are used to facilitate discussions and class participation. At the start of the term, all students are briefed about "Do's and Don'ts" (a.k.a. Slack usage etiquette). The guidelines include but are not limited to: 1) stay on topic, 2) search first and then ask if unable to find solutions online, 3) remember that nothing is private online, and 4) be careful with your tone and use respectful language.

4. Flip & Slack Model – Evaluations

This study involved two instructors, four teaching assistants, and 145 students. Three sets of data are collected to better understand students learning experience. To answer **RQ1** and **RQ2**, we conducted a Likert-scale based survey. It focuses on learning via two designs: Flip design and Slack design. To analyze the findings from survey, we collected the Slack data and explored the student interactions. To answer **RQ3**, **RQ4** and **RQ5**, we conducted the Likert-scale based survey on cognitive and affective skills. Both scales are on 5 points: Strongly agree (SA), Agree (A), Neutral (N), Disagree (D), Strongly disagree (SD). We also collected qualitative feedback during the survey process to conduct qualitative analysis on the model. In Tables 4, 5 and 6, column S is 'Skewness' and column **K** is 'Kurtosis'.

Table 4. *Survey Results on Impact of Flip-Slack Model on Learning Experience*

Learning	Question	SA	A	N	D	SD	S	K
	Flip design on learning							
General	The instructor has effectively organized each week's flipped classroom activities.	50%	46%	4%	0%	0%	-0.5	-0.7
PL	I learn more through flipped classroom method than the traditional method.	38%	27%	23%	12%	0%	-0.6	-0.7
PL	I can follow the lessons better after watching the videos.	50%	43%	6%	1%	0%	-1.5	4.1
PL	Learning through flipped classroom makes me anxious	11%	26%	24%	39%	0%	0.3	-0.9
CL	Interactions during in class hours are sufficient for learning.	22%	47%	18%	13%	0%	-0.5	-0.5
CL	My peer interactions in the classroom have strengthened.	18%	31%	32%	18%	0%	-0.2	-0.7
CL	My interactions with faculty in classroom have strengthened.	22%	37%	28%	13%	0%	-0.4	-0.5
	Slack design on learning							
General	Topics discussed are relevant to the course.	44%	49%	6%	1%	0%	-1.3	3.6
PL	Instructor gives us useful and immediate feedback in Slack.	55%	35%	10%	0%	0%	-1.4	2.8
PL	Slack helps me with reviewing the topics covered in class.	26%	46%	20%	7%	0%	-0.7	0.1
PL	Social interactions make me anxious.	7%	28%	27%	38%	0%	0.2	-0.9
CL	Slack allows me to share ways to solve the challenges in the course.	38%	43%	17%	2%	0%	-0.9	0.8
CL	Interactions in Slack complement the learning from classroom interactions.	19%	60%	17%	4%	0%	-0.9	1.7

CL	Using Slack, my interactions with my peers outside classroom have strengthened.	11%	19%	39%	31%	0%	0.4	-0.5
CL	Using Slack, my interactions with faculty outside classroom have strengthened	18%	41%	30%	11%	0%	-0.3	-0.4

Table 5. *Survey Results on Impact of Flip-Slack Model on Cognitive Skills*

Cognitivism	Question	SA	A	N	D	SA	S	K
Understanding	I am able to understand the concepts covered in the course.	47%	50%	3%	0%	0%	-0.3	-1
Remembering	I am able to remember (list, describe, identify, and answer questions) concepts.	33%	56%	9%	2%	0%	-0.6	0.8
Evaluating	I am able to debug the web layout using the concepts, using debugging tools.	36%	53%	9%	1%	1%	-1.1	2.9
Creating	I am able to create a web page using the concepts taught in the course.	41%	56%	3%	0%	0%	-0.1	-1
Analyzing	I am able to determine how web components are interrelated in a web.	41%	54%	5%	0%	0%	-0.2	-0.7
Applying	I am able to design a web layout using the components taught in the course.	40%	54%	5%	1%	0%	-0.6	0.8

4.1 Impact of flip & Slack Model on Learning

To evaluate *RQ1* and *RQ2*, we designed the survey questions to study the impact of flip on personalized learning (PL) and collaborative learning (CL). The summary of survey results is depicted in Table 4. Answer to *RQ1* from flip design survey: The data indicate moderately skewed for all questions except for “Following lessons” and more weight in tails (k=4.1). For all other questions, the values are within the range indicating reasonable sample size. The summative results of the survey on flip design suggest that flipped classroom concept is generally well received by students (65%) as compared to the traditional method, although a third of the students (37%) still feels anxious about such learning pedagogy. Students respond positively towards how the weekly class are organized (96%) and video-based learning (93%). We do observe that there are fewer interactions between the students (49%) and instructors (59%) in the classroom. Flipped classroom has a significant positive impact on the personalized learning and average positive impact on the collaborative learning. Peer interactions were not as successful due to the university-wide restriction on the maximum students allowed on campus.

Answer to *RQ2* from slack design survey: The data collected indicate moderately skewed for all questions except for “instructor feedback” and more weight in tails (k=2.8). For all other questions, the values are within the range indicating reasonable sample size. The survey results show that the use of Slack as a tool allows students to share and exchange solutions and knowledge in the weekly challenge (81%) and complements classroom learning (79%). Students can receive useful and quick feedback (90%) which are useful when they review on the topics covered in the course (72%). More importantly, students find the topics discussed on Slack are relevant to that of the course (93%). We do observe the similar pattern that there are fewer interactions between the students (30%) and instructors (59%) outside the classroom. Similarly, almost third of the students (35%) are anxious on Slack platform. To answer *RQ2*, slack interactions have a significant positive impact on the personalized learning and average positive impact on the collaborative learning. In our analysis, we observe that the peer interactions outside the classroom were not as successful as expected. Several factors may attribute to this behavior such are students are dependent on instructors on the answers for complex coding questions, or students of the same group tend to interact on the social media giving less room for peer collaborations.

4.2 Cognitive Skills Development

To evaluate **RQ3**, we designed the survey questions to study the impact of flip-slack on cognitive levels and learning outcomes. The summary of survey results is shown in Table 5. The data collected indicate moderately skewed for all questions except for "evaluating (s=-1.1)" and more weight in tails (k=2.9). For all other questions, the values are within the range indicating reasonable sample size. A large majority of the students indicate that they are able to understand, remember, evaluate, create, analyze, and apply web development concepts learned in the course. To answer **RQ3**, the flip-slack model has a significant positive impact on cognitive skills development.

4.3 Affective Skills Development

To evaluate **RQ4**, we designed the survey questions to study the impact of flip-slack on affective levels and student behavior. The summary of survey results is depicted in Table 6.

Table 6. *Survey Results on Impact of Flip-Slack Model on Affective Levels*

Affective levels	Question	SA	A	N	D	SA	S	K
Receiving	I read all the questions and answers discussed by my classmates.	17%	43%	19%	21%	0%	-0.4	-0.7
	I watch all the videos or content posted by the instructor.	50%	45%	5%	1%	0%	-0.9	0.9
Responding	I actively participate in class discussions.	13%	32%	35%	20%	0%	-0.1	-0.6
	I initiate ideas on the solutions to my classmates.	19%	31%	35%	15%	0%	-0.3	-0.5
Valuing	I support/debate with others' posts by like/dislike or appreciation or with replies (e.g. stickers).	22%	34%	29%	15%	0%	-0.4	-0.6
	I am committed to the course and respect my classmates.	51%	45%	4%	0%	0%	-0.5	-0.7
Organization	I know the Slack and classroom rules and ethics and follow them.	49%	46%	6%	0%	0%	-0.5	-0.6
	I plan and organize events/tasks systematically to solve problem.	23%	58%	15%	5%	0%	-0.7	0.6
	I am able to prioritize my time to meet the goals of the team.	39%	53%	6%	2%	0%	-0.8	1.3
Characterization	I am able to work on given problems independently.	31%	61%	5%	4%	0%	-1	2.1
	I am able to lead activities/ discussions in Slack and in class.	13%	26%	42%	19%	0%	0	-0.4
	I give objective problem solving methods.	15%	59%	23%	3%	0%	-0.7	1.5

The data collected indicate moderately skewed for all questions and more weight in tails for "time prioritization (k=1.3)", "independent work attitude (k=2.1)" and "provide objective solutions (k=1.5)". For all other questions, the values are within the range indicating reasonable sample size. The survey included questions to determine the affective outcome of active learning adopted in the course. While the students are committed to the course and respect their classmates (96%), only half show the appreciation with replies or indicators (56%) on Slack. While in class, less than half actively partake in class discussions (45%) and half propose solutions to their classmates (50%). Although students are receptive to postings by the instructors (95%), the percentage is much lesser for their peers (60%). On a positive note, high percentage of students indicate that they are aware of the ethical issues in Slack (95%). Students are able to plan and organize events and tasks systematically to solve the given problem (81%) and work towards the goals effectively (92%). Students are also able to work on the given problems independently (92%). To answer **RQ4**, the flip-slack model has an average positive impact on affective skills development. We performed a correlation analysis to further study the correlations between self-reflections on skills development and student course grades.

4.4 Correlations between the Skills Development and Grades

To answer **RQ5**, we use Pearson correlation coefficient scores between the skills ratings and student grades. Table 7 shows the correlations. For cognitive skills, Table 7 shows significant weak positive correlations with stronger p-values (<=0.05) for all levels except remembering. This may be because the students practice and prepare for exams whereas for the weekly classes, they are not putting efforts on remembering skills. For affective skills, Table 7 shows no significant evidence with higher p-values (<=0.05) for all affective levels. This indicates no relationship between grades and affective skills development. Recall from a previous analysis that students reported average scores on affective skills development through interactions. However, the grades of the students are higher indicating inconclusive results on affective skills. To answer **RQ5**, we observe a significant weak positive correlation between cognitive skills developed and grades whereas no significant evidence for affective skills developed. To deeply understand this observation, further investigation is required by performing content analysis on Slack messages.

Table 7. *Correlations between Self-assessed Skills Development and Course Grades*

Cognitive Skills	**Pearson r**	**p-value**		**Affective Skills**	**Pearson r**	**p-value**
Understanding	0.234	0.01*		Receiving	-0.108	0.27
Remembering	0.149	0.12		Responding	0.044	0.65
Evaluating	0.207	0.03*		Valuing	-0.012	0.90
Creating	0.241	0.01*		Organization	-0.052	0.59
Analyzing	0.217	0.02*		Characterization	0.164	0.09
Applying	0.17	0.05*				

5. Limitations & Future Work

This study is a first attempt to examine and evaluate the flip-slack model for designing and delivering computing courses. Several notable caveats are worth mentioning in our work. Firstly, a formal control group experiments would provide more persuading data with regards to the impact of the flip-slack model on the student learning experience. Secondly, students' attitudes and perceptions can influence their learning (Candeias et al., 2011; Shamsuddin et al., 2018). If students' attitudes towards and perceptions of a new form of learning (such as video-based learning, social media interactions, etc.) are positive, their learning experience is likely to be enhanced. Students that strongly prefer only instructor-led lectures may respond negatively towards this form of learning, and thus, their learning experience will suffer. Combining findings from a deeper investigation into students' attitudes and perceptions would lead to a more comprehensive understanding of factors that influence students' learning. Thirdly, from our survey, we also observed that students rated low for interactions and this factor may impact the learning experience of the students. Content analysis on the posts and peer interaction analysis using social network mining would help to understand the factors affecting the interactions. Finally, the affective skills are analyzed using the student self reflections and this can be a limitation of the survey approach. Emotion analysis using emoticons and mention analysis of acknowledgements provide substantial evidence for the insights into affective skills development.

6. Conclusion

This paper presents an integrated platform for cognitive and affective skills development. We share the design of an innovative flipped classroom model integrated with Slack and present the design principles applied to emphasize cognitive and affective skills development through the flip-slack model. In our case study, the analysis of the student survey responses reveals that flipped classroom and Slack interactions have a significant positive impact on the personalized learning and average positive impact on the collaborative learning. Further, it indicates that the flip-slack model has a significant positive impact on cognitive skills development and an average positive impact on affective skills development.

Acknowledgements

We would like to thank all teaching team members and students who participated in this research study.

References

Al-Ibrahim, A., and Al-Khalifa, H. S. (2014). Observing online discussions in educational social networks: A case study. *2014 International Conference on Web and Open Access to Learning (ICWOAL)*, 1-4.

Anderson, L. W., Krathwohl, D. R., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., et al. (2001). A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy. Longman Publishing.

Bergmann, J., and Sams, A. (2012). Flip Your Classroom: Reach every student in every class every day. Eugene, OR: International Society for Technology in Education.

Bishop, J. L., and Verleger, M. (2013). The flipped classroom: A survey of the research. *ASEE Annual Conference and Exposition, Conference Proceedings*.

Brown, G. A., and Atkins, M. (1996). Effective Teaching in Higher Education. Routledge.

Caceffo, R., Gama, G., and Azevedo, R. (2018). Exploring Active Learning Approaches to Computer Science Classes, *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*.

Candeias, A., Rebelo, N., and Oliveira, M. (2011). Student' Attitudes Toward Learning and School – Study of Exploratory Models about the Effects of Socio-demographics and Personal Attributes.

Davidson, N., and Major, C. (2014). Boundary Crossings: Cooperative Learning, Collaborative Learning, and Problem-Based Learning. *Journal on excellence in college teaching*, 25, 7-55.

Eppard, J., and Rochdi, A. (2017). A framework for flipped learning. *Proceedings of the 13th International Conference on Mobile Learning 2017*, 33-40.

Fassbinder, A., Botelho, T. G., Martins, R. J., and Barbosa, E. (2015). Applying flipped classroom and problem-based learning in a CS1 course. *2015 IEEE Frontiers in Education Conference (FIE)*, 1-7.

Ferreira, J. M. M. (2014). Flipped classrooms: From concept to reality using Google Apps. *11th International Conference on Remote Engineering and Virtual Instrumentation (REV)*, 204-208.

Foldnes, N. (2016). The flipped classroom and cooperative learning: Evidence from a randomised experiment. *Active Learning in Higher Education*, 17, 39 - 49.

Gehringer, E., and Peddycord, B. W. (2013). The inverted-lecture model: a case study in computer architecture. *Proceedings of the 44th ACM Technical Symposium on Computer Science Education*.

Goedhart, N., Westrhenen, N. B., Moser, C., and Zweekhorst, M. (2019). The flipped classroom: supporting a diverse group of students in their learning. *Learning Environments Research*, 22, 297-310.

Gokarn, M. N., Gottipati, S., and Shankararaman, V. (2019). Cognitive and Social Interaction Analysis in Graduate Discussion Forums. *Proceedings of the 2019 IEEE Frontiers in Education Conference (FIE)*, 1-8.

Grabinger, R. S., and Dunlap, J. C. (1995). Rich environments for active learning: A definition. *Association for Learning Technology Journal*, 3(2):5–34.

Hardaway, D. E., and Scamell, R. W. (2005). Use of technology-mediated learning instructional approach for teaching an introduction to information technology course. *JISE*, 16(2), 137-145.

Johnson, D. W., & Johnson, R. T. (2002). Cooperative learning and social interdependence theory. In Theory and research on small groups (pp. 9-35). Springer, Boston, MA.

Krathwohl, D. R., Bloom, B. S. and Masia, B. B. (1964). Taxonomy of educational objectives, Book II. Affective domain. New York, NY. David McKay Company, Inc.

Lockwood, K., and Esselstein, R. (2013). The inverted classroom and the CS curriculum. *Proceedings of the 44th ACM Technical Symposium on Computer Science Education*.

Maher, M., Latulipe, C., Lipford, H., and Rorrer, A. (2015). Flipped Classroom Strategies for CS Education. *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*.

Mentz, E., Walt, J. V., and Goosen, L. (2008). The effect of incorporating cooperative learning principles in pair programming for student teachers. *Computer Science Education*, 18, 247 - 260.

Moore, M. (1989). Three Types of Interaction. *American Journal of Distance Education*. 3. 1-7. 10.1080/08923648909526659.

Shamsuddin, M., Mahlan, S. B., Ul-Saufie, A. Z., Hussin, F., and Alias, F. A. (2018). An identification of factors influencing student's attitude and perception towards mathematics using factor analysis. *AIP Conference Proceedings*.

Shim, K. J., Gottipati, S. and Lau, Y. M. (2021) Integrated Framework for Developing Instructional Videos for Foundational Computing Courses. *PACIS 2021 Proceedings*. 217. https://aisel.aisnet.org/pacis2021/217

Skinner, B. F. (1948). Walden Two Indianapolis, IN: Hackett Publishing Company.

Thongmak, M. (2017). Flipping MIS Classroom by Peers: Gateway to Student's Engagement Intention. *Proceedings of the 26th International Conference on World Wide Web Companion*.

Topping, K. J., and Ehly, S. W. (1998). Peer-Assisted Learning. Lawrence Erlbaum Associates, ISBN9780805825022.

Watson, J. B., and Rayner, R. (1920). Conditioned emotional responses. *Journal of Experimental Psychology*, 3, 1-14.

Wu, C., Huang, Y., and Hwang, J. (2016). Review of affective computing in education/learning: Trends and challenges. *Br. J. Educ. Technol.*, 47, 1304-1323.

Yamamoto, J., and Ananou, S. (2015). Humanity in Digital Age: Cognitive, Social, Emotional, and Ethical Implications. *Contemporary Educational Technology*, 6(1), 1-18. https://doi.org/10.30935/cedtech/6136

Laboratory Study on ICAP Interventions for Interactive Activity: Investigation Based on Learning Performance

Shigen SHIMOJO[a*] & Yugo HAYASHI[b]
[a]*Graduate of Human Science, Ritsumeikan University, Japan*
[b]*College of Comprehensive Psychology, Ritsumeikan University, Japan*
*cp0013kr@ed.ritsumei.ac.jp

Abstract: Research on collaborative learning indicates that knowledge deepens through conversations. However, studies on providing prompts for support of collaborative learning have investigated support focusing on only interactive activities in the collaborative learning process much. The purpose of this study is to investigate the effect of providing interventions that respond to keywords based on the ICAP theory to facilitate collaborative learning process included subordinate processes based on the Interactive-Constructive-Active-Passive (ICAP) theory that learning science has focused on collaborative learning process. We make prompts responding to keywords to investigate whether learning performance is facilitated through them. Additionally, we set the control condition, random condition, and keyword condition. In keyword condition, systems detected keywords on ICAP, classified into interactive, constructive, active, and passive. As a result, the learning performance of learners in the keyword condition was higher than in the control condition. Consequently, prompts focusing on ICAP based on keywords are considered to have been effective in facilitating learning performance. This study also deals with issues and research contents that should be analyzed in the future.

Keywords: Collaborative learning, computer-supported collaborative learning (CSCL), ICAP framework

1. Introduction

In various fields such as learning science and cognitive science, both individual learning and collaborative learning have been studied. The effectiveness of collaborative learning is supported by the zone of proximal development theory, and it is understood that things that cannot be achieved as individuals can be achieved by collaborating with others (Vygotsky, 1980). Kupczynski, Mundy, Goswami, and Meling (2012) found that collaborative methods facilitate learning over individual methods in distance learning situations. Distance learning situation refers that learners cannot see each other, and learners learns in remote. However, it is difficult for teachers to provide appropriate guidance for each group, and the development of a support system independent of humans is needed.

Computer assisted instruction (CAI) was first investigated in research dealing with computer assistance. In recent years, there have been advancements in intelligent tutoring systems to provide adaptive feedback based on cognitive activity, with a focus on individual learning (Anderson, Corbett, Koedinger, & Pelletier, 1995). In contrast, computer-supported collaborative learning (CSCL) provides support using prompts that ask questions, and instructions and scripts that pre-specify learning activities to foster important processes in collaborative learning based on its theory. (Weinberger, Ertl, Fischer, & Mandl, 2005).

Many collaborative learning theories have been investigated, and it is necessary to study a system that supports collaborative learning based on them. In particular, in learning science research, there are many studies on the learning process (Meier, Spada, & Rummel, 2007). One of the studies examines the Interactive-Constructive-Active-Passive (ICAP) theory, which reveals that the learning process changes in stages, such as passive, active, constructive, and interactive (Chi & Wylie, 2014). It is suggested that it is important in learning science that learning process that learners contribute each other affect learning performance. Therefore, Our research group has investigated creating prompts based on cooperative and argumentative processes and presenting them randomly so far (Shimojo & Hayashi, in press). Consequently, it was found that learning is facilitated by randomly presenting prompts. However, the conversation process based on the ICAP theory was not observed sufficiently, and the change from passive to interactive was not fully clarified (Shimojo & Hayashi, 2020).

In this study, we focused on collaborative learning in distance situations. To facilitate a conversation process based on the ICAP theory, we detected keywords in conversation and investigated the effect of facilitating learning performance by presenting prompts accordingly. First, in this section, I will describe the ICAP theory that this study focuses on, introduce research to facilitate collaborative learning processes, and describe the purpose and hypothesis of this research based on these points. Second, I mainly explain how to conduct experiment and provide prompts. Third, I show the result of hypothesis testing and then discuss the result. Final, I summarize this paper.

1.1 ICAP Theory and Support for Collaborative Learning Based on ICAP Theory

In learning science, constructive interactions that share ideas and examine them are important (Miyake, 1986). Previous studies have focused on various collaborative learning processes, but this research focuses on the ICAP theory. The reason is that it is possible to grasp the depth of the process and to make a clear policy to support the passive voice to interact.

The ICAP theory (Chi & Wylie, 2014) classifies collaborative learning processes into interactive, constructive, active, and passive, and states that interactive is the deepest process where learning is most facilitated. Passive is where the person simply receives information on learning materials, active is where there is paraphrasing or repetition of learning materials, constructive process entails a deep understanding of learning materials, and interactive is an activity that further deepens understanding through interaction with others. The interactive process builds one's own opinion based on the opinion of the partner. These processes have a hierarchical structure. Therefore, in supporting collaborative learning, it is important to encourage learners to active, constructive, and interactive for facilitating interactive. The hypothesis that learning performance is facilitated as it becomes interactive, as mentioned earlier, is supported by the knowledge change process. In the passive process, learners simply store new information (store); in active, new information activates related prior knowledge and stores it in an integrated manner (integrate); in constructive, new information is integrated and inferred from active knowledge (infer); and finally, in the interactive process, each learner infers new knowledge from integrated/activated knowledge and from others (co-infer). That is, as the learner's activities become interactive, the knowledge deepens. Wiggins, Eddy, Grunspan, and Crowe (2017) investigate instructor support but find it difficult to engage learners interactively because it is difficult to facilitate the interactive process due to lack of time and human resources in the scene of large-scale collaborative learning with human support. Therefore, to facilitate interactive learning process in the order mentioned above, support for prompts based on context and timing using computer systems is required.

Accordingly, insufficient time and human resources are available in the case of purely human support, and thus, collaborative learning support using a computer is required. In the next section, we describe what kind of support method is useful for encouraging learners' interactive activities.

1.2 Facilitation of Collaborative Learning Process by Using Computer

Various studies in CSCL have been investigated to facilitate collaborative learning processes using computers. For example, Weinberger, Ertl, Fischer, & Mandl (2005) conducted a combined script and prompts, facilitating the processes at each stage, and demonstrating their usefulness not only text-based but also oral. Therefore, prompts are useful for facilitating a particular collaborative learning process. However, pre-ordered support such as conventional scripts has over-script problems, such as reduced motivation (Dillenbourg, 2002). Therefore, it is necessary to consider collaborative learning support based on the context and timing of collaborative learning.

Among such collaborative learning support, studies have examined the support methods for specific learning activities using agents. For example, Hayashi (2020) studied support using prompts and sensing technology. It has been clarified that learning performance is facilitated by detecting simple information such as keywords from conversations to facilitate collaborative learning processes and using prompts according to the context and timing to some extent. The effectiveness of contextual and timing-based prompts is to provide relevant support, and it has been found that providing such support fosters learning activity and facilitates learning performance (Walker, Rummel, & Koedinger, 2014). As a method of verifying the effect of support according to the context and timing, a method of comparing the control condition without support, random prompts condition and adaptive condition is used. However, as we have seen, it is insufficient to consider presenting prompts according to the context and timing from the viewpoint of the collaborative learning process. Therefore, to facilitate interactive, prompts that can foster a specific conversation activity is effective, and when investigating the prompt effect according to the context and timing, it is possible to compare it with random presentation.

Therefore, in this study, we examine whether it is more effective to provide related interventions considering context and timing than to randomly present interventions when facilitating interactive activities. As a method, prompts that respond to ICAP-based keywords are presented, and when a keyword related to the passive process is detected, prompts related to the active are presented to prompt the related process. After that, the learner’s conversation activity is facilitated for interactive in the same way. Since ICAP is an activity classified mainly based on utterances, it can be classified by utterances, and the keywords included in the utterances related to each activity are clear according to the task. Therefore, it is possible to detect keywords related to each activity to a certain extent, and it is possible to present prompts according to the context and timing.

1.3 Purpose and Hypothesis

The purpose of this study is to improve learning performance (pre-post test) by providing interventions that respond to keywords based on the ICAP theory to facilitate interactive process of a pair of learners during conversation. It is necessary to clarify whether it is facilitated. Therefore, to investigate the effect of presenting interventions that facilitate the interactive activities in response to keywords related to ICAP in oral dialog are provided, the following hypothesis is given and verified in the result section.

H1: Learners who are provided with ICAP interventions have better learning performance than learners who are not presented with ICAP interventions.

H2: Learners who are provided with ICAP interventions that respond to keywords related to ICAP are better at learning performance than learners who are provided with ICAP interventions randomly.

2. Method

2.1 Participants and Design

Participants in the experiment were 62 university students majoring in psychology. The average age was 19.21 years (*SD* = 1.08), and 18 were male and 44 were female. One pair of participants was excluded because of missing data.

In this study, we adopted a one-factor between-subject plan to examine two hypotheses. The inter-subject factor is a factor in the presentation of the prompt method, and three conditions were set, which were: control conditions, random conditions, and keyword conditions. Under the control conditions, only two people worked on the task without ICAP prompts on the screen. Under the random condition, prompts to facilitate ICAP was randomly presented on the screen while the two people were working on the task. In the keyword condition, as in the random condition, prompts to facilitate ICAP while working on the task by two people was presented on the screen. However, the presentation method detected keywords related to ICAP rather than random, and the prompt was presented based on the state of the ICAP. The specific system configuration is described in the section on the system structure below.

Participants in the experiment were randomly assigned to each of the three conditions and paired. Next, a test was used to confirm prior knowledge about causal attribution of whether there was any difference between the conditions. There was no difference between the conditions in the prior knowledge of cause attribution, which is a concept to be learned (F (2,57) = 2.11, p = 0.13, *partial* η^2 = 0.07). Also, they did not know about cause attribution so much.

2.2 Materials

The materials used in this study were learning texts on attribution theory, which is a concept to be learned, episodes for experimental tasks, and concept maps, which are tools used in experimental tasks.

2.2.1 Learning Text and Experimental Task

The learning texts were about attribution theory, a psychological theory about the causal attribution of success and failure. Specifically, the text explained the important concepts in the attribution of causes of success and failure: external/internal, stable/unstable, controllable/uncontrollable.

In this task, we infer the mental model of Peter, one of the characters, using the cause attribution of success and failure learned in this learning text. Peter is a student who feels anxious in the new semester, and learners was asked to read his episode and infer the mind of him by using causal attribution of success and failure. When making the inference, he created it using a concept map. This task was created based on Weinberger and Fischer (2006).

2.2.2 Concept Map

In this study, CmapTools (https://cmap.ihmc.us/) was used. Creating a concept map is effective for learning and is used for discussions in collaborative learning. During the task, it was necessary to share the concept map created by the individual with the collaborators and to create one concept map by two people. Figure 1 illustrates a screen capture of the collaboration.

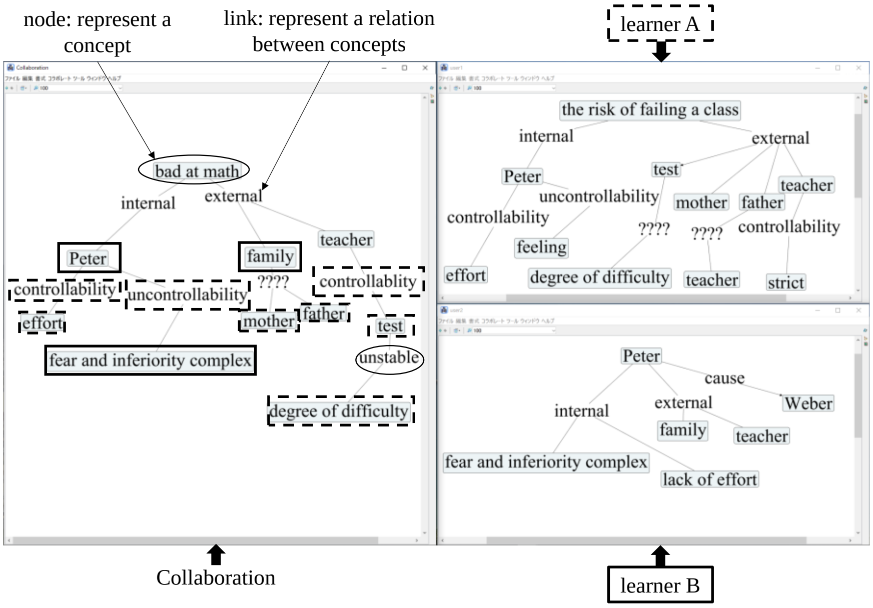


Figure 1. Screen Shot of Cooperation in this Task. Dotted line is for learner A and the solid line is for learner B. The ellipse is not included in either, and the rest are common.

2.3 Procedure

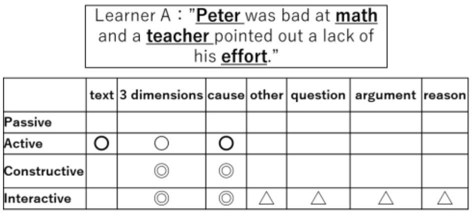
This experiment consists of three tests (check test, pre-test, post-test), a concept map explanation, a learning textbook, and the experimental task in individual and collaborative situations. Next, the specific procedure is explained.

First, a check test was performed to confirm whether the prior knowledge level was equivalent between the conditions. Next, the experimenter explained the concept map used in this task—that it consists of the relationship between concepts. After that, learners read and studied the prescribed textbook and answered the pretest. In this task, we asked them to infer Peter’s mental model individually using a concept map and then collaborate to create a concept map. Finally, learners answered the post-test. The only difference between the conditions was the collaboration phase of this task.

2.4 Structure of Prompt System

In this study, we used different systems for random conditions and keyword conditions. The installed prompts are of the same type in random and keyword conditions, created using C#, and TCP/IP communication was used for communication. First, the random condition presented prompts at random. In other words, the order of prompts was random.

In the keyword condition, a keyword list was first created to detect keywords related to ICAP. Figure 2 shows the example of keyword detection. The utterance of learner A was classified into active because “Peter” and “teacher” were in the list of “text” and “math” and “effort” were in the list of “cause”. Since the active process entails an utterance that repeats the information of the learning material, we used keywords such as effort, cause attribution and internal/external in the learning material. Also, we used keywords combined 3 dimensions (e.g., stable and unstable) and cause because constructive is a process that learners deepen text and infer cause based on 3 dimensions. Furthermore, we used keywords of collaboration in addition to constructive. The presentation method is to first recognize the keyword using voice recognition, detect the current process of ICAP, and if it is passive, present prompts to facilitate active, and if it becomes active, present prompts to facilitate constructive. However, when it became interactive, I presented prompts to facilitate the active process and proceeded as illustrated in Figure 3.



Learner A : "**Peter** was bad at **math** and a **teacher** pointed out a lack of his **effort**.”

	text	3 dimensions	cause	other	question	argument	reason
Passive							
Active	○	○	○				
Constructive		◎	◎				
Interactive		◎	◎	△	△	△	△

Figure 2. Example of Detection of Keywords and Relation between Keywords and ICAP. If there is one of three circles, active is. If there are two double circle, constructive is. If there are one of four triangles in addition to two of double circle, interactive is.

Active, constructive, and interactive prompts was made based on coding scheme of ICAP (Chi & Wiley, 2014). For example, we made “paraphrase the content of learning text” because active refers to paraphrasing leaning text. Also, we made “Explain why cause is classified in 3 dimensions” because constructive refers to deepening learning text. Furthermore, we made “explain one own idea by building on partner’s contribution” because interactive refers to build on partner’s contribution. Total number was 16 (active was 3, constructive was 5, and interactive was 8).

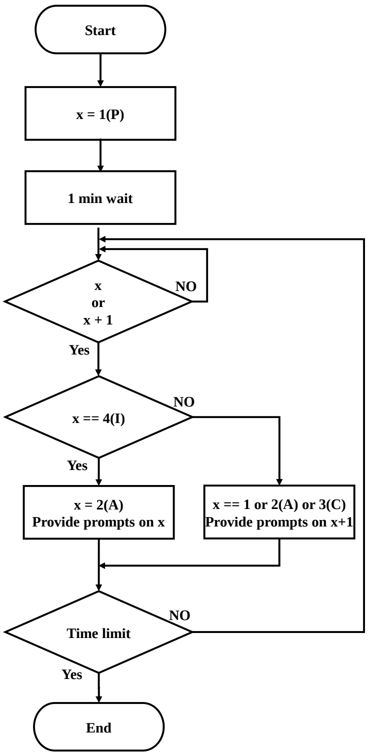


Figure 3. Flow Chart of Changes of ICAP States and Types of Facilitation Presented. Detection of keywords was conducted at all the time.

2.5 Dependent Variables

In this study, learning performance was used as the dependent variable. In the analysis, from the pre-test to the post-test, whether the score increased or not, it was divided, and the independence test was performed.

The rate of these test was rated by one coder. Rating was standardized as follows and conducted on a scale of 1 to 5 point is not correct/answer. 1 point is naïve but correct answer. 2 point is abstract answer based on learning text that include attribution theory. 3 point is answer based on learning text that include the causal attribution of success and failure in addition to that. 4 point is correct answer based on learning text that include 3 dimensions (internal-external, stable-unstable, controllability-uncontrollability) in addition to those. 5 point is specific answers deepened the learning text that include specific explanation in addition to those. The specific explanation is description using example and deep content based on the experimental task and learning text. Accordingly, learning performance measures the depth of understanding and knowledge on learning text and learners was needed to engage in interactive. Also, we calculated Krippendorff’s alpha coefficient between first coder and second coder to investigate the reliability. As a result, the coder’s matching rate was 0.48. Therefore, that the coding was reasonably reliable and the coding of first coder was adopted.

In addition to learning performance, we analyzed the ratio of new node exploratory (number of new nodes / total number of nodes in cooperation). The new node is related to the task. If added node was not related to the task, we didn’t count. This makes it possible to easily confirm whether new knowledge or viewpoints are generated through the task.

3. Result

3.1 Concept Maps

We analyzed the concept map created in this task and confirm whether the performance of the task was affected between the conditions. Therefore, we calculated the ratio of new nodes (number of new nodes / total number of nodes in cooperation) and performed a one-factor analysis of variance. For example, the number of new nodes of learner A in Figure 2 is 4, and the total number of nodes in cooperation is 10, so the ratio of new nodes is 40%. The number of new nodes of learner B is 5 and the ratio of new nodes is 50%. Figure 4 summarizes the average values of the new node ratios calculated in this way for each condition in Task 2. Thus, a difference was observed between the conditions (*F* (2,57) = 3.74, *p* <.05, *partial η2* = 0.12). As a result of multiple comparisons by Shaffer's method, it was found that the ratio of new nodes for learners of keyword conditions was higher than that of control conditions (*p* <.05). It is suggested that the effect of interventions that facilitate interactive process based on keywords was reflected in the task.

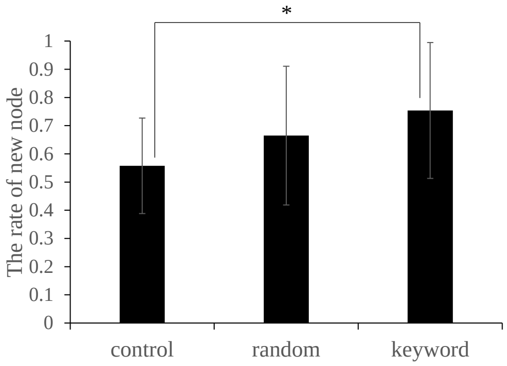


Figure 4. Comparison between the Conditions of the Ratio of New Nodes in the Concept Map in this Task 2. Error bars are the standard deviation and * represents *p* < .05.

3.2 Learning Performance

In this section, we analyze the learning performance to verify H1 and H2. Specifically, from the pre-test to the post-test, the score was classified into whether the score increased (with an increase) or not (without an increase), and Fisher's test was performed. Two factors (prompts presentation method) and learning performance were confirmed. Table 1 presents the cross-tabulation table. There was a significant difference between conditions with prompts and without prompts (*p* <.001). Next, as a result of the residual analysis, it was found that under the control conditions, there was a significantly lower increase and no increase was significantly higher. Under the keyword condition, an increase was significantly high, and no increase was significantly low. Furthermore, it was found that there was a significant increase in keyword conditions rather than in control conditions.

From the above, it can be said that H1 and H2 were partially supported because the keyword condition facilitated the learning performance more than the control condition.

Table 1. *Cross Table of Control Conditions, Random Conditions, and Keyword Conditions in Learning Performance*

	Without an increase	With an increase	Total
control condition	10 3.77 **	10 -3.77 **	20
random condition	3 -0.89	17 0.89	20
keyword condition	0 -2.88 **	20 2.88 **	20
Total	13	47	60

4. Discussion

In this study, H1 that the learner who presented the intervention facilitated the learning performance more than the learner who did not present the intervention, and H2 that the keyword condition facilitated the learning performance more than the random condition. From the results, H1 and H2 were partially supported. Therefore, it was suggested that the intervention that facilitates interactive process based on context and timing using keywords is effective to some extent. However, between the learner's learning performance presented at random and the learner presented with prompts according to the context and timing, there was no evident difference.

Previous studies examining support for collaborative learning have revealed that more contextual and timing-based support is beneficial for learning (Walker, Rummel, & Koedinger, 2014). However, it has not been sufficiently examined whether developing prompts based on the collaborative learning process in this study and determining the presentation method facilitates learning. In previous studies, unlike this study, the learner's problem-solving process was modeled, and prompts were presented based on it (Walker, Rummel, & Koedinger, 2014), and the learner's line-of-sight pattern information was used to identify the learner's state. Studies have been conducted to detect the state and give feedback accordingly (D'Mello, Olney, Williams, & Hays, 2012). Therefore, in the future, it is necessary to respond to the context and timing more accurately, that is, to detect the learner's state and provide adaptive support (e.g., prompts) accordingly. However, it is highly useful to see the effect of the interventions responding to the ICAP-based keywords in this study.

We examined why there was no difference between the keyword and random conditions. Therefore, we used the log of keywords detected by the system under keyword conditions. The average number of changes from passive or active to interactive was 3.00 (*SD* = 1.95), the most change from passive to interactive was 6, and the least was 0. Also, we conducted an exact binominal test. Table 2 shows the number of learners who were from passive to interactive in order by using keyword that the prompt system detected. As a result, there was a significant difference between learners who were not from passive to interactive in order and learners who were from passive to interactive in order (*p* < .05). This result indicated that ICAP intervention based on context and timing was effective. However, some learners was not from passive to interactive.

Table 2. *The Number of Learners Who Were from Passive to Interactive in Order*

	Not from passive to interactive	From passive to interactive	Total
Keyword condition	4	16	20

The fewest pairs transitioned from the passive to the constructive process but stopped in the constructive stage from the beginning to the end. That is, it was suggested that some pairs in the keyword condition did not reach the interactive stage, and there was a problem in promoting changes from constructive to interactive. When the pairs with the most changes from passive to interactive were divided into learners, it was found that the learners who made statements about the interactive process were biased. Therefore, future support methods can be seen from the perspectives of cooperation and individuals. First, from the perspective of cooperation, when the system is stopped in the constructive stage, it is necessary to provide more enforceable support in the future, in addition to the prompts that facilitate the utterance of the ICAP used this time. For example, for the learner to refute the opinions of others, display the counter-argument input window and prompt them to input. By doing so, it is thought that conversations that are biased toward one's own opinions shift to an interactive state, in which opinions are constructed based on each other's ideas. From an individual point of view, it is necessary to separate the prompt presented according to each learner and the prompt presented according to the pair when presenting the prompt. This is because if the utterances related to the interactive process are biased, it is necessary to eliminate the bias and prevent it in learning outcomes in pairs. In the future, we will build a prompt presentation model from these two perspectives and examine how the interactive process is improved compared to this keyword condition.

In this study, we focused on learning performance, but as mentioned earlier, it is necessary to focus on the collaborative learning process and quantitatively analyze its depth. The study conducted this time is a rough analysis because the change was examined from the keywords captured by the system. Therefore, it is necessary to conduct a detailed analysis of how much the prompt that responds to the keyword could be facilitated the interactive process. Furthermore, it is necessary to confirm the effect of the support method by checking whether the ratio of the interactive process differs between the conditions. Therefore, in the future, we will focus on the collaborative learning process and examine its depth of the learning process.

5. Conclusion

It has been clarified that the learning process and related knowledge have deepened in the collaborative learning process. However, previous studies on prompt presentation in collaborative learning have not sufficiently examined the support method focusing on the deepening of the collaborative learning process of pairs of learners. In addition, when providing support, it is necessary to provide more relevant support in consideration of context and timing. Therefore, this study aimed to examine the effect of interventions to change collaborative learning process from the learner's utterance information based on the ICAP theory, focusing on the deepening of the learning process. To verify the effect of this intervention, we created prompts that respond to the utterance keywords defined based on each process defined as per the ICAP theory. When they were presented, we observed that learning performance was facilitated. In the experiment, we prepared a control condition that does not present prompts, a random condition that presents prompts randomly, and a keyword condition that prompts from passive to interactive in stages and conducted a laboratory experiment. As a result, it was found that the keyword presentation condition that facilitates interactive process based on conversation information facilitates learning performance more than the condition that does not present the prompts. In contrast, there was no difference in the conditions presented at random. Therefore, there was an effect of prompting from the passive to interactive process in stages to some extent. However, in this study, it is not possible to analyze in-depth the extent to which the change in the learning process can be facilitated. Therefore, a detailed analysis focusing on the collaborative learning process is necessary in future studies.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP20H04299.

References

Shimojo, S. & Hayashi, Y. (2020). Prompting Learner-Learner Collaborative Learning for Deeper Interaction: Conversational Analysis Based on the ICAP Framework. *Proceedings of the 28th International Conference on Computers in Education*, 177–182.

Shimojo, S. & Hayashi, Y. (in press). Facilitating explanation activities using a concept map in collaborative learning: 7 Focusing on coordination and argumentation process. Cognitive Studies.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive Tutors: Lessons Learned. *The Journal of the Learned Sciences*, *4*(2), 167-207.

Chi, M. T. H., & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist*, *49*(4), 219-243.

Dillenbourg, P. (2002). Over-scripting CSCL: The risks of blending collaborative learning with instructional design. In P. A. Kirschner (Ed.), *Three worlds of CSCL. Can we support CSCL?*. Heerlen: Open Universiteit Nederland.

D'Mello, S., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human Computer Studies*, *70*(5), 377-398.

Hayashi, Y. (2020). Gaze awareness and metacognitive suggestions by a pedagogical conversational agent : an experimental investigation on interventions to support collaborative learning process and performance. *International Journal of Computer-Supported Collaborative Learning*, *15*, 469-498.

Kupczynski, L., Mundy, M. A., Goswami, J., & Meling, V. (2012). Cooperative Learning in Distance Learning: a Mixed Methods Study. *International Journal of Instruction*, *5*(2), 81-90.

Meier, A., Spada, H., & Rummel, N. (2007). A rating scheme for assessing the quality of computer-supported collaboration processes. *International Journal of Computer-Supported Collaborative Learning*, *2*(1), 63-86.

Miyake, N. (1986). Constructive Interaction and the Iterative Process of Understanding. *Cognitive Science*, *10*(2), 151-177.

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard university press.

Walker, E., Rummel, N., & Koedinger, K. R. (2014). Adaptive intelligent support to improve peer tutoring in algebra. *International Journal of Artificial Intelligence in Education*, *24*(1), 33-61.

Weinberger, A., Ertl, B., Fischer, F., & Mandl, H. (2005). Epistemic and social scripts in computer-supported collaborative learning, *Instructional Science*, *33*(1), 1-30.

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, *46*(1), 71-65.

Wiggins, B. L., Eddy, S. L., Grunspan, D. Z., & Crowe, A. J. (2017). The ICAP Active Learning Framework Predicts the Learning Gains Observed in Intensely Active Classroom Experiences. *AERA Open*, *3*(2),1-14.

Designing Support for Productive Social Interaction and Knowledge Co-Construction in Collaborative Annotation

Xinran ZHU*, Hong SHUI & Bodong CHEN
Department of Curriculum and Instruction, University of Minnesota, U.S.A
*zhu00323@umn.edu

Abstract: This paper introduces a generic scaffolding framework of participation roles that was co-designed by instructors and researchers to support collaborative learning activities in online classes. Informed by the CSCL literature, the framework specifies three participation roles – *facilitator*, *synthesizer*, and *summarizer* – that play distinct roles in each week's collaborative activities. Using a web annotation tool named Hypothes.is, we piloted the framework in a fully online undergraduate course in Fall 2020. To examine how the framework facilitated social interaction and knowledge co-construction in the class, we conducted social network analysis and content analysis on students' annotation data generated from their engagement with 18 readings. Results indicated the participation roles were enacted properly to a great extent and knowledge co-construction was facilitated when role-takers made high-level contributions. This study has practical implications for online teaching and collaborative learning.

Keywords: Collaborative annotation, participation roles, social interaction, knowledge co-construction

1. Introduction

The COVID-19 pandemic has forced instructors around the globe to seek ways to engage learners in disciplinary learning and peer interaction. While some instructors focused on replicating models of face-to-face (F2F) teaching, others took this opportunity to explore affordances provided by web technologies for new models of instruction. Take seminar courses for example, where critical reading and classroom dialogues are often important means to achieve higher-order competencies such as critical thinking, communication, and collaboration. While it is often believed that in-depth dialogues could only take place in a F2F setting, we can also argue that F2F communication in a fixed amount of time poses serious constraints for classroom dialogues, limiting the amount of speaking opportunities and favoring learners who are more outspoken in a F2F environment. In contrast, web technologies, such as web annotation and video conferencing, offer opportunities for learners to participate in different ways than being in a F2F setting.

This paper reports on a pilot study conducted at a large public university in the US in Fall 2020, when the university campus was shut down due to COVID-19 and many instructors pivoted to online instruction. To engage students in reading and discussing course materials in three liberal-arts classes, we collaborated with instructors to integrate a web annotation technology, *Hypothes.is*, in their online teaching. While the extant literature has investigated various usages of web annotation tools in classrooms (Zhu et al., 2020), studies that incorporate computer-supported collaborative learning (CSCL) ideas in the design of social annotation activities remain rare. In this study, we designed a scaffolding framework comprising three predefined participation roles for learners to participate in weekly collaborative reading and annotation activities. This study advances CSCL and online learning research by generating a design framework of collaborative annotation and testing it in online courses. For the remainder of this paper, we first introduce key perspectives informing this study. We then describe the study context and research methods. After reporting the main findings, we discuss implications and future directions.

2. Related Literature

2.1 Using Web Annotation in Education

Annotation, be it online or paper-based, is an important part of human cognition. Making annotations is a highly developed activity, one that represents an important part of reading, writing, and scholarship (Marshall, 1997). For example, readers annotate printed books as a routine part of their engagement with the materials, with annotations serving a multitude of functions: procedural signals, placemarks, an in-situ way of working problems, interpretive activity, a visible trace of a reader's attention, and so on (Marshall, 1997; O'Hara & Sellen, 1997). While annotations are generally considered marginal, peripheral, and secondary, in-depth analysis of annotations in used books has revealed their added value to the "primary" content and their influence on later book users (Liu, 2005).

Web annotation is a genre of information technology that allows a user to annotate information in a shared web document and hereby anchor a discussion to the annotated information. Similar to annotations in paper-based documents, web annotations are extra pieces of information associated with existing, "first-order" web resources (Haslhofer et al., 2011). According to a systematic literature review, web annotation has been used across different education levels to help students process domain-specific knowledge, promote argumentation and inquiry, improve literacy skills, support instructor and peer assessment, and connect online learning spaces (Zhu et al., 2020). While some use cases of web annotation involve students reading and annotating in groups, there remains untapped potential in web annotation to promote collaborative learning through sophisticated CSCL designs.

2.2 Designing Participation Roles in CSCL

CSCL has a long-standing interest in designing sophisticated social configurations, such as participation roles and classroom discourse, for collaborative learning. This interest is grounded in CSCL's recognition of social interaction as an important factor of learning along with cognitive factors such as knowledge construction (Dillenbourg et al., 2009). In CSCL research, roles have been recognized as a fundamental aspect of group dynamics essential for collaborative knowledge construction (Heinimäki et al., 2020; Ouyang & Chang, 2019). Prior work has explicated two types of roles: *emerging roles* that participants develop naturally and spontaneously in their collaborative learning; and *scripted/assigned roles* that are usually pre-defined by the instructor or instructional designer to facilitate collaboration (Kollar et al., 2006; Strijbos & Weinberger, 2010).

The notion of emerging roles highlights learners' agency in structuring and regulating their collaborative processes. Emerging roles are dynamic over time in relation to learners' cognitive and social engagement (Strijbos & Weinberger, 2010). Reflecting CSCL's interest in scaffolding collaboration, participation roles are also designed to meet instructional goals (Strijbos & Weinberger, 2010). These roles can be designed in response to learner characteristics and curriculum objectives. One premise of this work is that students can meaningfully engage with content and with each other by assuming their assigned roles. Prior work has demonstrated by carefully assigning roles – either contented-oriented roles (e.g., summarizer) or activity-oriented roles (e.g., project planner) (Wise et al., 2012) – learners could harness productive interdependence to reach higher levels of knowledge construction, learner responsibility, and collaboration (Strijbos & Weinberger, 2010).

2.3 The Present Study

The study aims to support collaborative web annotation in college classrooms by designing sophisticated participation roles. Following a co-design approach, we worked closely with three instructors from a large public university in the U.S. to design a generic scaffolding framework for collaborative annotation activities and supported each of them to implement the framework in their classes, with course-specific customization. This study was conducted when the university pivoted to online/distance teaching in Fall 2020 and instructors were looking for ways to meet their teaching needs. At that time, the instructors were participating in a college-level pilot of a web annotation tool named Hypothes.is that was integrated in Canvas to support social reading and annotation among students.

In the design phase, we designed a generic scaffolding framework comprising three scripted participation roles based on the CSCL literature (Strijbos & Weinberger, 2010; Wise et al., 2012). These roles are: a *facilitator* responsible for stimulating conversations by finding connections, seeking clarifications, and encouraging their peers to consistently tag their annotations for an entire week; a *synthesizer* who synthesizes the initial ideas, highlights agreement/disagreement, and suggests directions of further discussions in the middle of the week; and a *summarizer* who summarizes group conversations at the end of the week for the whole class.

In the implementation phase, each instructor further customized the participation roles based on the class they taught. Figure 1 presents an example design from one class, which is the focus of this paper. In this class, each week, the instructor assigned readings and the participation roles to students. The students annotated the course readings and interacted with each other by replying to the annotations. The *facilitator* was responsible for catalyzing productive conversations throughout the week. Under facilitation, students negotiated the meaning of key terms from different perspectives. Figure 2 shows an example interface of the activity (Student B was the facilitator). The *synthesizer* collected students' different perspectives and reflected on their initial thoughts in the middle of the week before their class meeting on Zoom. During the class meeting, students discussed their annotations to address problems of understanding based on multiple perspectives. After the class discussion, the *summarizer* summarized the entire week's activities before each student wrote their individual reflection.

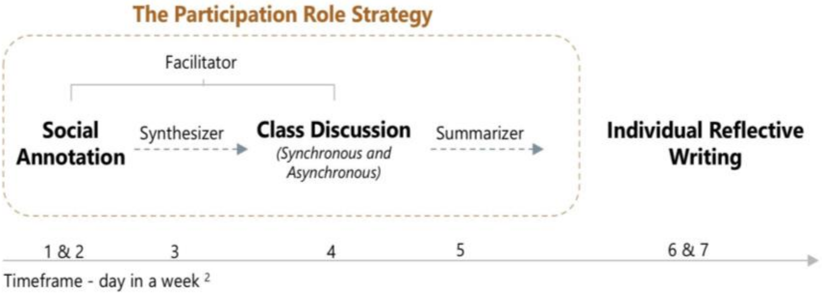
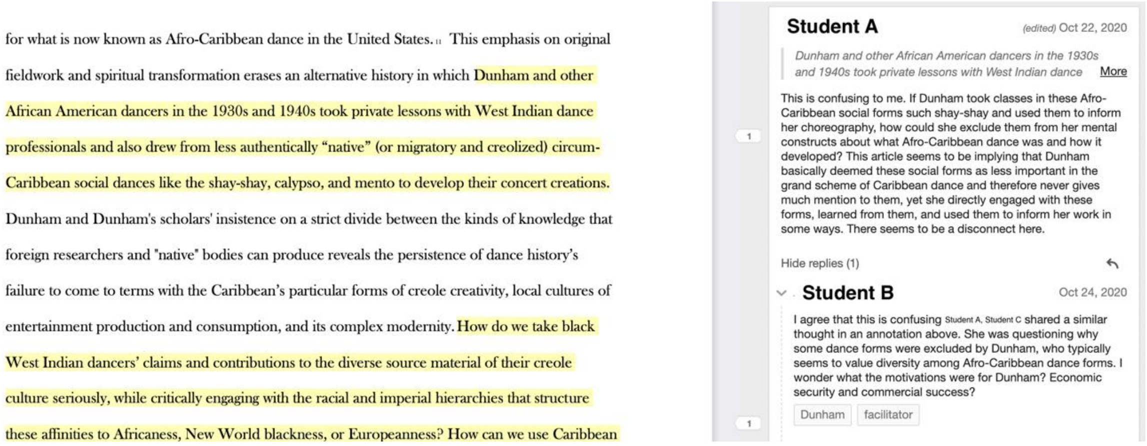

The Participation Role Strategy
Facilitator
Social Annotation
Synthesizer
Class Discussion
(Synchronous and Asynchronous)
Summarizer
Individual Reflective Writing
1 & 2
3
4
5
6 & 7
Timeframe - day in a week 2

Figure 1. The Scaffolding Framework of Participation Roles.



for what is now known as Afro-Caribbean dance in the United States.11 This emphasis on original fieldwork and spiritual transformation erases an alternative history in which Dunham and other African American dancers in the 1930s and 1940s took private lessons with West Indian dance professionals and also drew from less authentically "native" (or migratory and creolized) circum-Caribbean social dances like the shay-shay, calypso, and mento to develop their concert creations. Dunham and Dunham's scholars' insistence on a strict divide between the kinds of knowledge that foreign researchers and "native" bodies can produce reveals the persistence of dance history's failure to come to terms with the Caribbean's particular forms of creole creativity, local cultures of entertainment production and consumption, and its complex modernity. How do we take black West Indian dancers' claims and contributions to the diverse source material of their creole culture seriously, while critically engaging with the racial and imperial hierarchies that structure these affinities to Africaness, New World blackness, or Europeanness? How can we use Caribbean

Student A (edited) Oct 22, 2020

> *Dunham and other African American dancers in the 1930s and 1940s took private lessons with West Indian dance* More

This is confusing to me. If Dunham took classes in these Afro-Caribbean social forms such shay-shay and used them to inform her choreography, how could she exclude them from her mental constructs about what Afro-Caribbean dance was and how it developed? This article seems to be implying that Dunham basically deemed these social forms as less important in the grand scheme of Caribbean dance and therefore never gives much mention to them, yet she directly engaged with these forms, learned from them, and used them to inform her work in some ways. There seems to be a disconnect here.

Student B Oct 24, 2020

I agree that this is confusing Student A, Student C shared a similar thought in an annotation above. She was questioning why some dance forms were excluded by Dunham, who typically seems to value diversity among Afro-Caribbean dance forms. I wonder what the motivations were for Dunham? Economic security and commercial success?

Figure 2. An Example of the Collaborative Annotation Interface.

We proposed the following research questions to guide our investigation of the enacted participation roles strategy:

1. How did the activity design facilitate social interaction? In particular: What were the participation patterns for different participation roles? What were the participation patterns for the whole class and how were they related to patterns of participation roles?
2. How did the activity design facilitate knowledge co-construction? In particular: How were the levels of knowledge co-construction reflected in contributions made by different participation roles? How were the levels of knowledge co-construction reflected in contributions made by the whole class each week and how were they related to knowledge co-construction levels of participation roles?

3. Methods

3.1 Context and Participants

This study was conducted in a fully online undergraduate course at a large public university taught by one instructor and a teaching assistant in Fall 2020. In this liberal-arts class, students (n=13) were engaged in reading course materials, participating in weekly online meetings, and writing reflective essays. For the reading tasks, students were required to read 1-2 readings each week, post annotations on Hypothes.is, and reply to each other's annotations. Following the designed scaffolding framework, the instructor assigned the participation roles – i.e., *facilitator*, *synthesizer*, and *summarizer* – to three students each week from Week 1 to Week 11. Students rotated across weeks and had the opportunity to assume different roles.

3.2 Data Source

The main data source included 482 Hypothes.is annotations and 492 replies created by students in 18 readings across 11 weeks.

3.3 Data Analysis

To answer our research questions concerning the social and cognitive aspects of collaborative annotation, we analyzed social interaction and knowledge construction from a socio-cognitive perspective.

3.3.1 Social Network Analysis

To answer the first research question, Social Network Analysis (SNA) was applied to analyze participation patterns in the collaborative annotation activity. SNA as a methodology is interested in capturing and characterizing social positions, structures, and processes. It can capture the structure of a complete network as well as an individual's positions and behaviors in a network. For example, Dowel and Poquet (2021) used SNA measures such as *degree centrality* and *positional dominance* to capture learners' positions in massive open online courses (MOOCs). Such SNA measures could be further combined with other analytical methods (such as content analysis) to examine online communication. In this study, we conducted both whole-network and ego-network analysis to examine the role takers' positions and interaction patterns and their association with features of the full network. We first constructed interaction networks for the whole class, treating each student as a *node* and their interaction/reply events as *edges*; this network was temporal (sliced by week), directed (following the direction of replies), and weighted (based on the number of ties in a particular week). Network measures including *degree centralization*, *density*, *reciprocity*, and *transitivity* were calculated to characterize interaction patterns among students. From the full networks, we also extracted one-step ego networks for individual students and calculated ego-network measures including ego-network *size*, *centrality*, and *constraint* (Burt, 1992) to characterize each student's local situation. Explanations of these network analysis techniques are beyond the scope of this paper and can be found in texts such as Carolan (2014).

3.3.2 *Content Analysis*

To answer the second research question, we conducted content analysis using a coding scheme we developed for the social annotation context based on Gunawardena's Interaction Analysis Model (IAM) (1997) and Onrubia & Engel's model of collaborative knowledge construction (2009).

IAM divides knowledge construction into five phases: (1) Sharing and comparing information; (2) Discovering and exploration of dissonance or inconsistency among ideas, concepts or statements; (3) Negotiation of meaning/co-construction of knowledge; (4) Testing and modification if proposed synthesis or co-construction; and (5) Agreement statements/application of newly constructed meaning. One limitation of this model is that the highest level is rarely achieved. Research suggests that the scope of higher levels of knowledge construction needs to be reconsidered (Lucas et al., 2014). Besides, the discussion of dissonance as described in the IAM model may not be a necessary condition for higher levels of knowledge construction in certain contexts (Lucas et al., 2014). To address the limitations, we also referenced Onrubia & Engel (2009)'s model which identified four phases of collaborative knowledge construction: (1) Initiation; (2) Exploration; (3) Negotiation; and (4) Co-construction. This model is similar to IAM in terms of the typology of collaborative knowledge construction processes but merges IAM's 4th and 5th phases into one single phase (Lucas et al., 2014). In our study, we did not directly use the four-phase model developed by Onrubia & Engel (2009) because it is designed for collaborative writing activities that require student groups to negotiate dissonance and build consensus in order to to generate a shared writing document. In our context of collaborative reading and social annotation, consensus building was less of a concern as more emphasis was placed on the sensemaking and negotiation of ideas in the readings. Therefore, we developed a revised interaction analysis model (see Table 1). We adopted the levels from the four-phase model by Onrubia & Engel (2009) but revised the indicators by addressing the tasks of collaborative reading and social annotation.

Using this coding scheme, two researchers independently coded student annotation data from Week 1, compared the coding results (Cohen's *kappa* = .90), and addressed disagreements through discussion. After establishing a shared understanding of the coding scheme, each researcher coded half of the remaining data.

To investigate the extent to which participation roles facilitated knowledge co-construction, we first calculated knowledge co-construction levels for roles takers across readings to describe their level of knowledge co-construction in general. Then we zoomed into each reading to count the number of posts in each level contributed by non-role participants. By revealing the knowledge co-construction level of role takers and investigating their associations with the whole class's annotations, we explored how role takers were linked with the knowledge co-construction levels of their peers.

Table 1. *Revised IAM of Collaborative Annotation*

Level	Definition	Examples
Level-1: Initiation	a) Share initial understandings b) Ask questions and share resources without elaboration or critical examination	"Does this sound similar to what is happening in our society today? "
Level-2: Exploration	a) Elaborate on the texts b) Provide additional evidence/information to an argument without critical examination c) Make connections without critical examination	"Do you think this definition of social dance is accurate? What examples of social dance do we see today? How do these dances impact culture?"
Level-3: Negotiation	a) Response to questions through critical reasoning b) Negotiate disagreement c) Connect readings with critical reasoning d) Synthesize meanings e) Create new supporting statements by building on a previous conversation	"This also reminded me of the readings ... This approach to viewing performances seems desirable because it's often nice to just be able to watch a piece for the art that it is, but it is also important not to settle into this mindset and block out the intentions and messages behind a staged performance as well."
Level-4: Co-construction	a) Reach a consensus on a previous question b) Apply the knowledge or way of thinking gained through the activity c) Make a metacognitive statement illustrating their learning outcome	"… before this class began, I only thought of the first description when I considered diaspora. I viewed it as a lonely and isolating thing where people are forced from their homelands and lose all connection with their culture. However, these articles are broadening my view and allowing me to appreciate the connective power of diaspora, which I think is perfectly alluded to in this quote."

4. Results

4.1 *How Did the Activity Design Facilitate Social Interaction?*

4.1.1 Node-level Measures for Role Takers and Non-Role Takers

The mean and standard deviation of the SNA measures for different role takers suggested that the role takers, especially the *facilitators* and *synthesizers*, varied in the SNA measures which implies they may take different strategies when completing their tasks. The Analysis of variance (ANOVA) to determine if there are statistical differences between mean SNA measures among the four groups (non-role takers, *facilitators*, *synthesizers* and *summarizers*) suggested that the differences are significant in in-degree ($F(3, 211) = 3.48$, $p < .05$), out-degree ($F(3, 211) = 21.92$, $p < .05$), betweenness ($F(3, 211) = 5.67$, $p < .05$), positional dominance ($F(3, 211) = 6.16$, $p < .05$), and ego size ($F(3, 211) = 4.56$, $p < .05$). The constraint was not significantly different among groups.

A post hoc comparison using the *Tukey* test was also conducted to further examine the differences between each group (see Table 2). The results revealed that the *facilitators* were significantly different from non-role takers in all SNA measures except constraint. Also, the *facilitators* were significantly different from the *summarizers* in betweenness. The *synthesizers* were significantly different from non-role takers and *summarizers* in out-degree. The *summarizers* did not show significant difference with non-role takers in all measures.

The results aligned with the design that *facilitators* tended to facilitate the social interaction by sending out more replies and reaching out to more peers, resulting in receiving more replies and being influential in the collaborative annotation activities. The *synthesizers* also participated more than non-role takers in terms of the numbers of posts they sent out (out-degree), but not as much as the *facilitators* did in facilitating the interaction since they tended to focus more on synthesizing the readings and annotations. The *summarizers* participated the same as non-role takers since their responsibility was to write the weekly summary independently after class meetings.

Table 2. *Pairwise Comparisons among Groups*

		Mean Differences (A-B)					
Group A	Group B	In Degree	Out Degree	Betweenness	Constraint	Dominance	Ego Size
Facilitator	Synthesizer	0.11	0.03	5.21	-0.07	0.09	0.76
	Summarizers	0.07	0.14	10.13*	-0.05	0.14	1.24
	Non-role	0.11*	0.13*	9.75*	-0.08	0.16*	1.37*
Synthesizer	Facilitator	-0.11	-0.03	-5.21	0.07	-0.09	-0.76
	Summarizers	-0.04	0.11*	4.92	0.02	0.05	0.47
	Non-role	0.00	0.10*	4.55	-0.01	0.07	0.61
Summarizers	Facilitator	-0.07	-0.14*	-10.13*	0.05	-0.14	-1.24
	Synthesizer	0.04	-0.11*	-4.92	-0.02	-0.05	-0.47
	Non-role	0.04	-0.01	-0.38	-0.03	0.02	0.13

Note. * indicates the mean difference is significant at the .05 level.

4.1.2 Network-Level SNA Measures Across 11 Weeks

Whole-network SNA was conducted for each reading across 11 weeks. The results do not show discernible trends across weeks. For example, Reading 3a has the highest transitivity but relatively lower scores in the other network measures, while Reading 4 has relatively high scores among all four network measures. Readings 3b, 5, and 9b show a big drop in the SNA measures in comparison with previous weeks/readings, especially in degree centralization, transitivity, and reciprocity.

To further explore the explanations of the variance of the social interaction patterns across weeks/readings, a Pearson correlation test between *facilitators* and *synthesizers*' node-level measures and the network-level measures was conducted. As shown in Table 3, the network-level measures (except reciprocity) are significantly correlated with *facilitators* and *synthesizers*' node-level measures to some extent. It is worth noting that the positional dominance and the ego size of *facilitators* and all measures of *synthesizers* are linked to the network transitivity, e.g., the *synthesizers'* sending out more annotations (higher out-degree) is linked to a more transactive network (higher transitivity), meaning that students are more likely to develop different perspectives by interacting with multiple peers. The results suggested that these role takers' participation is associated with the interaction patterns for the whole class. Hence, when different role takers took different strategies to play their roles and interact with peers, it may lead to the variance of interaction patterns across the whole class.

Table 3. *Pearson Correlations between Facilitators and Synthesizers' Node-level Measures and Network-level Measures*

		Density	Reciprocity	Transitivity	Centralization
Facilitator	In-degree	0.32	0.34	0.33	0.68*
	Out-degree	0.83*	-0.02	0.47	0.74*
	Betweenness	0.57*	0.26	0.12	0.53*
	Constraint	-0.45	-0.19	-0.04	-0.28
	Dominance	0.43	0.31	0.50*	0.86*
	Ego size	0.60*	0.18	0.49*	0.67
Synthesizer	In-degree	0.57*	-0.18	0.64*	0.65*
	Out-degree	0.77*	0.13	0.62*	0.53*
	Betweenness	0.64*	-0.08	0.58*	0.45
	Constraint	-0.48	0.27	-0.55*	-0.05
	Dominance	0.48	0.27	0.57*	0.81*
	Ego size	0.75*	-0.13	0.77*	0.53*

Note. * indicates the correlation is significant at the .05 level.

4.2 How Did the Activity Design Facilitate Knowledge Co-Construction?

4.2.1 Knowledge Co-Construction Levels of Participation Roles in General

According to Table 4, the great numbers of Level-2 and Level-3 posts of the *facilitators* revealed that the *facilitators* generally asked questions or provided answers with elaboration, examples, critical reasoning, etc. to launch and advance the discussion. Yet the large standard deviation also suggested that the knowledge construction level varied across the *facilitators* in different weeks; some *facilitators* posted more Level-1 posts that consisted of only general questions or links to additional resources. Similarly, the *synthesizers*' posts were also mostly classified into Level-2 and Level-3 posts (83 out of 93 posts). It was partly because the scripted role of the *synthesizer* requested them to synthesize the initial ideas, highlight agreement and disagreement, and suggest directions for further conversations.

The *summarizers* on average contributed much less annotations. The results were in line with the design, i.e., the *summarizer* focused on the class discussion during Zoom meetings and composed a summary that connected synchronous Zoom discussions with asynchronous web annotations.

Table 4. *Mean and Standard Deviation of Participation Roles in Four Levels*

	Level-1	Level-2	Level-3	Level-4
Facilitator	0.88 (1.65)	2.24 (1.35)	3.24 (2.17)	0.18 (0.39)
Synthesizer	0.62 (0.81)	2.00 (0.89)	3.06 (1.73)	0.12 (0.50)
Summarizer	0.29 (0.47)	2.06 (1.09)	1.29 (1.05)	0.06 (0.24)

4.2.2 Knowledge Co-Construction Levels of Non-role Participants

According to the results in Figure 3, non-role rakers demonstrated comparatively higher knowledge co-construction levels in Readings 3a, 04, 6a, 8 and 11b. These five readings showed a similar growing trend in the frequency of levels, i.e., with very few Level-1 posts, a moderate quantity of Level-2 posts, and a great number of Level-3 posts. In addition, non-role takers in Readings 3a, 8 and 11b even contributed Level-4 posts that were rare in this dataset.

By contrast, non-role takers in Readings 2b, 3b, 6b, and 10b contributed more Level-1 posts and less Level-3 posts compared with the other readings, displaying a lower level of knowledge co-construction.

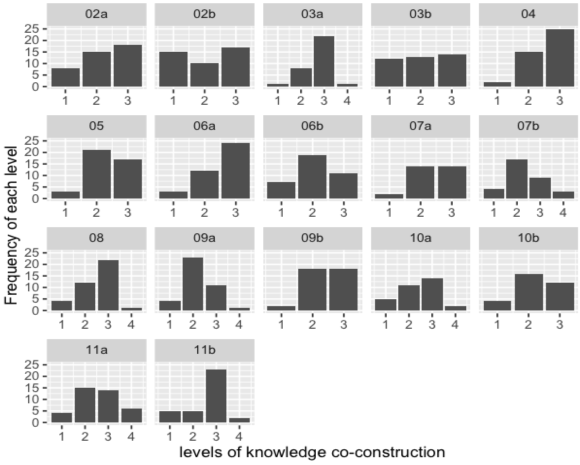


Figure 3. Level Frequency by Non-role Participants in Each Week. The first week was not measured because the instructor and TA played the participation roles as a demonstration.

4.2.4 *The Relationship between the Contributions Made by Role-takers and Non-role Takers*

Table 5. *The Percentage of Posts Contributed by the Role-takers in Each Knowledge Co-construction Level and the Average Knowledge Co-construction Levels of Non-roles*

		Readings															
		02a	02b	03a	03b	04	05	06a	06b	07a	07b	08	09a	09b	10a	10b	11a
Role	Level-1	6%	21%	5%	27%	0	0	0	43%	9%	32%	16%	13%	8%	8%	8%	0
	Level-2	56%	43%	45%	27%	30%	40%	21%	29%	35%	42%	32%	50%	58%	33%	42%	47%
	Level-3	38%	36%	35%	45%	65%	60%	79%	29%	57%	26%	53%	38%	33%	42%	50%	53%
	Level-4	0	0	15%	0	4%	0	0	0	0	0	0	0	0	17%	0	0
Non-role	average	2.24	2.05	2.72	2.05	2.55	2.34	2.54	2.11	2.40	2.33	2.51	2.21	2.42	2.41	2.18	2.56

Note. Reading 11b was excluded due to the absence of synthesizer and a very small number of posts contributed by the facilitator and summarizer. Percentages may not add up to 100 due to rounding.

Table 5 shows the percentage of posts contributed by the role takers in each knowledge construction level and the mean knowledge co-construction scores of non-role takers in each reading. In weeks when role takers posted more higher-level posts, the knowledge construction level from non-role takers tended to be high too. For example, in Readings 3a and 04 where 95% and 100% of the role takers' posts respectively were higher than Level-1, the knowledge construction levels for non-role takes were among the highest. Take the *facilitator* for example. In Reading 3a, when the *facilitator* sent out seven posts that all were above Level-2, they attracted 2 Level-2 replies and 6 Level-3 replies. To illustrate the details, below is one conversation thread demonstrating how this *facilitator* proposed specific questions to invite their peer to go deeper in this discussion.

> [Student 110]: Cultural syncretism means the blending of cultures to form something new. This can be in the form of religious practices, architecture, philosophy, recreation, food, etc. I think this back and forth Dunham was experiencing throughout her career is understandable. Was she in search of a right and a wrong answer? Or was she struggling to see how cultural syncretism preserved culture while simultaneously creating something new and different.
>
> > [Facilitator]: Student 110, this is a good thought and a new word for me, too. Student 105 student 114 talked about diaspora and assimilation a few paragraphs above. How do you think diaspora and syncretism relate, or maybe they do not relate at all? Do you think one is more beneficial than the other for preserving the culture?
> >
> > > [Student 110]: In general terms, I interpreted diaspora meaning this shift of cultures due to movement, and the intertwining of different cultures. I think syncretism focuses more on the combination of religious beliefs and an

> "interfaith". I don't know if one is better than the other, there always seems to be two sides to the story. In my opinion, I think the creation and development of new cultures is beautiful, but I am also someone who likes to hold onto tradition.

In contrast, Reading 6b's *facilitator* only received two replies to their five Level-1 annotations. Because the annotations made by this *facilitator* lacked in specificity and elicitation, e.g., "Here is a video of zapateado…," they failed to elicit contributions from others or to deepen the discussion.

5. Discussion

Inspired by the use of scripted roles to facilitate collaboration in CSCL, we worked with instructors to co-design a generic scaffolding framework for collaborative annotation activities by assigning three participation roles: *facilitator*, *synthesizer*, and *summarizer*. We piloted the design in a fully online undergraduate course and answered two research questions via SNA and content analysis on student annotation data.

The first question asked: How did the activity design facilitate social interaction? The ANOVA post-hoc pairwise comparison revealed that there was a significant difference between *facilitators* and non-role takers in annotation activities. It indicated that *facilitators* were most active in fostering the social interaction of the class by initiating the conversation through proposing questions, providing answers to puzzles, sharing information, etc. Besides *facilitators*, *synthesizers* also facilitated the social interaction by connecting readings and annotations to further the negotiation. *Summarizers*' participation was similar to non-role takers. These results indicated that to a great extent the designed activity was enacted by students properly.

The second question was: How did the activity design facilitate knowledge co-construction? Generally, *facilitators* and *synthesizers* held higher knowledge co-construction levels than *summarizers*. Examining weekly contributions of participation roles and non-role participants indicated that the knowledge co-construction level of role takers was associated with the level of their peers. For instance, when a *facilitator* sent out high level posts, they were more likely to receive replies and trigger a negotiation among peers. The reason might be that the high-level posts -- which featured elaboration, connection, critical reasoning, and application -- provided more directions for peers to engage in the conversation.

This paper's contribution to the CSCL and online learning literature is three-fold. First, we proposed a scaffolding framework for collaborative annotation that is applicable to many college-level classes. This framework builds on prior frameworks developed for online discussion forums (e.g., Wise et al., 2012) and extends CSCL ideas to support collaborative reading and annotation. Second, we developed a revised Interaction Analysis Model for collaborative annotation that is more appropriate for analysis of student discussions "anchored" in web documents. Finally, results of data analysis have shown promise of the designed scaffolding framework for facilitating productive collaborative annotation in the study context. In particular, the *facilitators* and *synthesizers* played roles in deepening collaborative annotation.

These findings have practical implications for online and hybrid classes. First, assigning students to different participation roles, such as *facilitators* and *synthesizers*, is worth considering in classes that involve asynchronous communication. Even though *facilitators* may participate in different manners, when they make high-level contributions that ask well-reasoned questions or make important connections between ideas, the quality of student discussion could be enhanced. Second, the study also implied that students need support to assume different participation roles. Indeed, students are not always natural collaborators and need to make intentional efforts to become better collaborators (Borge & White, 2016). The instructor needs to provide careful scaffolding and detailed guidelines for students to take various roles.

This paper only reports preliminary findings from a series of studies that attempt to facilitate collaborative annotation in college classrooms. There are several future directions for this work. First, we plan to deepen the analyses presented in this paper by incorporating advanced network modeling to examine the effects of social and cognitive factors on peer interaction. We are also in the process of analyzing two other classes that implemented the scaffolding framework. We plan to compare results among these classes to identify commonalities and differences. Finally, we are working on designing new tools for students to assume these participation roles more effectively. These efforts are all geared towards discovering means to promote new genres of collaborative learning that are supported by CSCL theories, digital tools, and instructional models that are tested in real-world settings.

References

Borge, M., & White, B. (2016). Toward the development of socio-metacognitive expertise: An approach to developing collaborative competence. *Cognition and Instruction*, *34*(4), 323–360. https://doi.org/10.1080/07370008.2016.1215722

Burt, R. S. (1992). Structural holes. In *structural holes*. Harvard University Press.

Carolan, B. V. (2014). Social network analysis and education: Theory, methods & applications. Sage Publications.

Dillenbourg, P., Järvelä, S., & Fischer, F. (2009). The evolution of research on computer-supported collaborative learning. In Technology-enhanced learning (pp. 3-19). Springer, Dordrecht.

Dowell, N. M., & Poquet, O. (2021). SCIP: Combining group communication and interpersonal positioning to identify emergent roles in scaled digital environments. *Computers in Human Behavior*, *119*, 106709.

Gunawardena, C. N., Lowe, C. A., & Anderson, T. (1997). Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of educational computing research*, *17*(4), 397-431.

Haslhofer, B., Simon, R., Sanderson, R., & Van de Sompel, H. (2011, September). The open annotation collaboration (OAC) model. In 2011 Workshop on Multimedia on the Web (pp. 5-9). IEEE.

Hou, H.-T., Chang, K.-E., & Sung, Y.-T. (2009). Using blogs as a professional development tool for teachers: Analysis of interaction behavioral patterns. Interactive Learning Environments, 17(4), 325–340.

Heinimäki, O. P., Volet, S., & Vauras, M. (2020). Core and Activity-Specific Functional Participatory Roles in Collaborative Science Learning. *Frontline Learning Research*, *8*(2), 65-89.

Kollar, I., Fischer, F., & Hesse, F. W. (2006). Collaboration scripts–a conceptual analysis. *Educational Psychology Review*, *18*(2), 159-185.

Liu, Z. (2005). Reading behavior in the digital environment: Changes in reading behavior over the past ten years. *Journal of documentation*.

Lucas, M., Gunawardena, C., & Moreira, A. (2014). Assessing social construction of knowledge online: A critique of the interaction analysis model. Computers in Human Behavior, 30, 574-582.

Marshall, C. C. (1997, July). Annotation: from paper books to the digital library. In *Proceedings of the second ACM international conference on Digital libraries* (pp. 131-140).

O'hara, K., & Sellen, A. (1997, March). A comparison of reading paper and on-line documents. In Proceedings of the ACM SIGCHI Conference on Human factors in computing systems (pp. 335-342).

Onrubia, J., & Engel, A. (2009). Strategies for collaborative writing and phases of knowledge construction in CSCL environments. *Computers & Education*, *53*(4), 1256-1265.

Ouyang, F., & Chang, Y. H. (2019). The relationships between social participatory roles and cognitive engagement levels in online discussions. *British Journal of Educational Technology*, *50*(3), 1396-1414.

Schellens, T., Van Keer, H., De Wever, B., & Valcke, M. (2007). Scripting by assigning roles: Does it improve knowledge construction in asynchronous discussion groups? Computer-Supported Collaborative Learning, 2, 225–246.

Strijbos, J. W., & Weinberger, A. (2010). Emerging and scripted roles in computer-supported collaborative learning. *Computers in Human Behavior*, *26*(4), 491-494.

Wise, A. F., Saghafian, M., & Padmanabhan, P. (2012). Towards more precise design guidance: Specifying and testing the functions of assigned student roles in online discussions. *Educational Technology Research and Development*, *60*(1), 55-82

Zhu, X., Chen, B., Avadhanam, R.M., Shui, H. and Zhang, R.Z. (2020), "Reading and connecting: using social annotation in online classes", *Information and Learning Sciences*, Vol. 121 No. 5/6, pp. 261-271.

Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

A Measure to Cultivate Engaged Peer Assessors: A Validation Study on its Efficacy

Yu-Hsin LIU[a], Kristine LIU[b] & Fu-Yun YU[c*]
[a]*Department of Civil Engineering, National Chi Nan University, Taiwan*
[b]*Medill School of Journalism, Media, Integrated Marketing Communications, Northwestern University, USA*
[c]*Institute of Education, National Cheng Kung University, Taiwan*
**fuyun.ncku@gmail.com*

Abstract: Despite the generally positive learning effects of peer assessment, undesirable behaviors exhibited during the process have been reported (e.g., peer assessors engaging at a superficial level, or giving biased judgements). With reference to related literature and based on non-participant observation of student assessors' behavior in classrooms and document analysis of past student assessors' work, a measure consisting of four variables was devised to serve two purposes: on the passive side, to alleviate such reported hindrances; on the active side, to engage peer assessors in sensible, prudent ratings. The devised measure quantifies the performance of student assessors based on the scores they give to their peers' performance/work as compared to that of the teacher/expert on each assessment criterion and two other noteworthy variables (i.e., fine discrimination ability, completion rate). A validation study which involved two classes of university sophomores was conducted. The students presented individual projects while participating in assessing their peers' performance via an online system, which only differ in whether the devised measure was in use (the experimental group, N = 53) or not (the comparison group, N = 47) for the two respective class. The statistically different results in peer assessors' performance between the two treatment groups, t (98) = 8.97 < .001, attested the efficacy of the devised measure in encouraging higher quality peer assessments.

Keywords: Expert ratings, performance assessment, peer ratings, online peer assessment, the quality of peer assessment

1. Introduction

Peer assessment has been a subject of investigation since Topping's (1988) highly cited review paper. Studies investigating the potential of peer assessment for the support of the teaching and learning process have mushroomed over the years. Empirical studies accumulated over the last three decades have generally confirmed the positive effects of peer assessment for promoting academic performance (Double, McGrane, & Hopfenbeck, 2020; Li, Xiong, Hunter, & Guo, 2019), deeper learning (Li, Bialo, Xiong, & Hunter, 2020; Sluijsmans et al., 2002), core 21st century skills development (Yu & Wu, 2016), and positive social-affective outcomes (vanGennip, Segers, & Tillema, 2009).

In light of the many affordances of networked technology, online peer assessment has been an emerging area that continues to attract increasing attention from academics and educators (Kulkarni et al. 2015). Currently, many online peer assessment platforms are available in the market to support the assessment of students' produced work/performance and the effort and time invested by the involved collaborative group-members. While supportive evidence on learning gains from these developed systems has been reported, some undesirable signs and behaviors exhibited during peer assessment have been noted. Examples of these aspects include peer assessors engaging at a superficial level and giving biased, unfair judgements (Adachi, Tai & Dawson, 2018; Liu & Carless, 2006; Yu & Sung, 2019). Hence, devising efficacious designs to help alleviate these challenges of online peer assessment is a worthwhile and important endeavor.

In this work, a peer assessment measure is proposed. The peer assessment measure, in the form of an equation, quantifies the quality of the assessors' assessment and consists of four variables. A validation study was conducted to provide preliminary data on the efficacy of the devised peer assessment measure.

2. An Online Peer Assessment Measure to Promote Engaged Process

Knowing that the relative accuracy of peer and teacher/expert ratings is a major concern of educators (Li et al., 2015), the correlation between peer and teacher/expert ratings takes center stage in the devised equation. In addition, non-participant observation of student assessors' behavior exhibited during the peer assessment process in classrooms and document analysis of past student assessors' ratings of their peer work/performance from the previous academic year were employed to identify other noteworthy variables — differentiation of peer work/performance along each criterion and completion rate. In short, the equation devised to quantify the quality of each student assessor's performance consists of four variables: $Y = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4$, where w denotes the respective weighting associated with each variable. Each of the four x variables is explained below.

Considering that peer assessment criteria may be subjective (e.g., organization, logic, clarity, appeal, interest, visual design, conciseness, etc.) and objective (e.g. within the limit of time or page) by nature, variable x_1 deals with all criteria of a subjective nature whereas variable x_2 deals with the objective type. Both variables x_1 and x_2 are obtained by calculating the correlation between a student assessor's score with that of the teacher/expert on each respective criterion. With reference to Li et al. (2015) meta-analysis which found the estimated average Pearson correlation (i.e., r) between peer and teacher ratings to be moderately strong (i.e., r = .63 to be exact), the calculated r is translated to a score by referring to Tables 1 and 2 for subjective and objective criteria, respectively. Generally, the student assessor and teacher/expert's low r corresponds to a low score (with calculated r below 0 being given a scoring of 0) whereas higher r corresponds to a higher score.

Compared to subjective criteria (x_1), objective criteria (x_2) is expected to have a clear-cut benchmark to compare against. Thus, it would demand a higher correlation between the score of the student assessor and that of the teacher/expert than subjective criteria. As shown in Tables 1 and 2, when the calculated r equals to or higher than 0.83, it is given a full score of 100 for subjective criteria. As for objective criteria, the calculated r would need to equal to or higher than 0.90 to be given a full score of 100. Also, the translated scores of all criterion are summed and averaged for subjective and objective criteria, respectively.

Table 1. x_1 *Scores Corresponding to Given r Values for Subjective Criteria*

Calculated r value	Scoring
[0.83, 1.00]	100
(0.63, 0.83)	100*(0.17 + r)
0.63	80
[0.00, 0.63]	100*(0.17 + r)
(0.0, -1.0)	0

Table 2. x_2 *Scores Corresponding to Given r Values for Objective Criteria*

Calculated r value	Scoring
[0.90, 1.00]	100
[0.00, 0.90]	100*(0.10 + r)
(0.0, -1.0)	0

For the x_3 component, we adopt the concept of item discrimination as stressed by psychometrics to denote the ability of a test item to discriminate amongst examinees (Amedahe & Asamoah-Gyimah, 2016). The spread and range of student assessors' scoring on each assessment criteria are considered in our devised equation. To this aim, standard deviation (denotes *sd*), which connotes the spread of scores one gives to their peers, is used to see if student assessors possess the intended fine differentiation on each criterion. Conceptually speaking, high *sd* would represent high differentiation ability while concentration of scoring around a rating scale is characterized by low *sd*. Nonetheless, to account for undesired possible polarized scoring, which would also result in high *sd*, x_3 is proposed to be calculated as listed in Table 3. Here, M = maximum possible value of *sd*, which will change depending on the number of points in a scale. M is calculated as: (the number of points in a scale - 1)/2. For example, for a 7-point scale, M = (7-1)/2 = 3.

Table 3. x_3 *Scores Corresponding to Calculated sd Value*

Calculated *sd* value	Scoring
[0.00, (1/3)M)	100*(3/M)**sd*
[(1/3)M, (2/3)M]	100
((2/3)M, M]	100*(3/M)(M – *sd*)

The last variable attributing to the quality of student assessors' performance, x_4, considers peer assessment task completion. It is calculated by counting the total number of peer assessment forms individual student assessor submitted against the expected number of forms to be submitted. The score the student assessor receives is proportional to the rate of which s/he completes the assigned peer assessment. Student assessors with a 100% completion rate are granted a full score on this metric.

3. A Validation Study on the Proposed Measure Efficacy

3.1 The Participants, Context, and Online Learning System

Two classes of college sophomores (*N* = 100) taking a required three-credit course (i.e., Transportation Engineering) from a university located in central Taiwan participated in this validation study. As part of the course requirement, the participants were asked to make an oral presentation with PowerPoint on their chosen topic. They were also asked to assess their classmates' presentation. The participants were informed that both their performance on PowerPoint presentation and peer assessment activity accounted for 20% of their final grade.

An online system to support online peer assessment activity developed by the authors was extended by embedding the devised equation to be activated by the instructor. Each of the participants would rate their peers' presentation on each of the devised criteria of the instructor's choice on their choice of personal devices.

3.2 Research Design, Data Analysis, and Test Results

To test the efficacy of the devised equation, a validation study involving two intact classes was conducted. Both participating classes used the same online system for the peer assessment activity. The only difference between the two group lies in whether the equation function is activated (the experimental group, *N* = 53) or not activated (the comparison group, *N* = 47). The participants of the experimental group were simply told that with reference related literature, a measure to objectively quantify their performance at peer assessment activity was in place. However, the participants had no knowledge as to what variables were involved or how their assessment scores were calculated.

For this validation study, five criteria were devised for the peer assessment activity. They are the overall visual design of PowerPoint slides, oral communication (e.g., fluency, clarity, logic), content (e.g., organization), control of presentation time, and presence and appearance. The participants were directed to rate their peers' presentation on a 7-point Likert scale on each of the five criteria. Minimal instruction was provided for the four subjective criteria as the participants' judgements may vary contingent on individual liking. Nonetheless, for objective assessment, a scoring instruction was developed for the objective criterion (i.e., the 'control of presentation time' criterion of this study) (see Figure 1 for the score to be given for the assessment of control of presentation time.

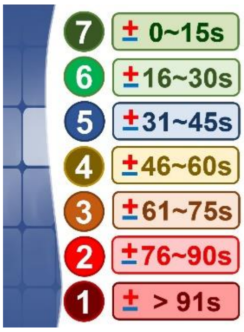


Figure 1. Scoring instruction given and posted for objective criteria — control of presentation time

In this validation study, the instructor of this course played the role of the expert. The expert respective scores given to individual presentation was compared to that given by each student assessor in a criterion-by-criterion fashion and then averaged before being translating to x_1. For the validation test, the weighting for the four variables were arbitrarily set at: 0.4, 0.2, 0.2, and 0.2 for x_1, x_2, x_3, and x_4, respectively, where each of the four subjective criteria accounts for 10% of presentation score.

Descriptive statistics of the two participating classes in peer assessment performance are listed in Table 4. Further independent sample *t*-test performed found a statistical difference between the two treatment groups, $t\ (98) = 8.97 < .001$. In other words, integrating the devised measure in the online peer assessment system helps promote higher quality peer assessment.

Table 4. *Descriptive Statistics of Peer Assessment Performance in the Two Treatment Groups*

Treatments Statistics	The comparison Group (*N* = 47)	The experimental Group[*](*N* = 53)
Mean	62.74	84.98
Standard Deviation	15.94	6.27

[*]The devised measure integrated in the adopted online peer assessment system

4. Conclusions

The learning effects associated with peer assessment have generally been confirmed positively. Yet, a couple of undesirable signs and behaviors exhibited during online peer assessment have been noted. With reference to related literature and based on non-participant observation and document analysis of student assessors' behavior and ratings in the previous semester, a measure consisting of four variables was devised to engage peer assessors in prudent ratings while alleviating implicit personal bias from affecting peer evaluation scores. As evidence by the validation study, the devised measure is efficacious in promoting overall better peer assessment performance.

References

Adachi, C., Tai, J. H-M, & Dawson, P. (2018). Academics' perceptions of the benefits and challenges of self and peer assessment in higher education, *Assessment & Evaluation in Higher Education, 43*(2), 294-306.

Amedahe, F. K., & Asamoah-Gyimah, K. (2016). *Introduction to measurement and evaluation* (7th ed.). Cape Coast: Hampton Press.

Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review, 32*, 481–509.

Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D. &, S. R. Klemmer (2015). Peer and Self Assessment in Massive Online Classes. In H. Plattner, C. Meinel, and L. Leifer (eds) *Design thinking research* (pp. 131–168). Cham: Springer

Li, H., Bialo, J., Xiong, Y., & Hunter, C. V. (2020). Effects of peer assessment on students' non-cognitive outcomes. *American Educational Research Association Annual Meeting*. San Francisco.

Li, H., Xiong, Y., Hunter, C. V., & Guo, X. (2019). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education 45*(1),1-19.

Li, H., Xiong, Y., Zang, X., Kornhaber, M. L., K., Lyu, Y., Chung, K. S., & Suen, H. K. (2015). Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education, 41*(2), 245-264.

Liu, N.-F. & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education, 11*(3), 279–290.

Sluijsmans, D. M. A., Brand-Gruwel, S., van Merriënboer, J. J. G., & Bastiaens, T. J. (2002). The training of peer assessment skills to promote the development of reflection skills in teacher education. *Studies in Educational Evaluation, 29*(1), 23–42.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249-276.

van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: The influence of interpersonal variables and structural features. *Educational Research Review, 4*(1), 41–54.

Yu, F. Y. & Sung, H. S. (2019). Online targeting behavior of peer-assessors under identity-revealed, created, and concealed modes. *Educational Technology and Society, 22*(1), 15-27

Yu, F. Y. & Wu, C. P. (2016). Predictive effects of the quality of online peer-feedback provided and received on primary school students' quality of question-generation. *Educational Technology & Society*, *19*(3), 234-246.

STEM and Non-STEM Students' Perception towards Work Environment and Career Prospect

Priscilla MOSES*, Tiny Chiu Yuen TEY & Phaik Kin CHEAH
Universiti Tunku Abdul Rahman, Malaysia
*priscilla@utar.edu.my

Abstract: The Malaysian education system has undergone transformation to reinstate the significance of science, technology, engineering, and mathematics (STEM). All students at the upper secondary schools are given the chance to learn STEM subjects so that they are equipped with the knowledge and skills needed for STEM careers despite streams of study at schools. Considering the importance of the new streaming system that divides students into STEM and non-STEM stream, it is not clear how students from both streams of studies perceive work environment and prospect of STEM careers. Hence, the purpose of this study was to examine if there is a significant difference between STEM and non-STEM stream secondary school students' perceptions towards STEM work environment and career prospect. Independent-samples t-tests was used to analyse data collected from 157 students from the East Coast of Malaysia. Though the effects of the results were small, this study found that students from different streams of study perceive working environment and career prospect in STEM differently. This research finding would offer insights to the authorities and policy makers to improve the new streaming system in the Malaysia to better prepare students for STEM careers.

Keywords: STEM, non-STEM, perceived work environment, career prospect

1. Introduction

The current demand for STEM talents surpasses its supply in STEM education (Shahali et al., 2017). Numerous vacancies in STEM jobs have created a worldwide concern because lacking STEM human capital for the workforce will threaten a country's development and the global economy at large (Ali et al., 2021). In view of this issue, the Malaysian government has initiated a national STEM action plan to inform the public, schools, and particularly the students about awareness, importance, and career opportunities in STEM.

Many activities, complementary programmes, and out-of-school activities have been implemented to offer students the necessary exposure to STEM beyond academic (Shahali et al., 2017). Through these initiatives, students are expected to be more motivated and are more likely to develop interest to pursue STEM in future (Shahali et al., 2017). Outreach programmes and collaborations with STEM agencies and organisations are among the most popular activities that allow students to have a glance at the actual work environment and foresee the prospects of STEM careers (Shahali et al., 2017).

According to Zhang et al. (2020), a supportive work environment such as having an appropriate reward system can generate positive job performance at the workplace, whereas an unfavourable work environment such as having heavy workload and unrealistic expectations would intimidate the driving force for work dynamic. Besides, Wan et al. (2014) noted that individuals who have possess high value over a career indicate high likeliness to pursue and commit to the targeted industry for its better career prospects. Given the importance of work environment and career prospect in career decision, having understood students' perceived work environment and career prospect in STEM would be able to offer better insights on how to inculcate a desirable and competitive STEM workforce. However, currently it is not clear how STEM and non-STEM students perceive work environment and prospect of STEM careers. Therefore, this study is an initial attempt to test whether there is a significant difference between STEM and non-STEM stream secondary school students' perceptions towards work environment and career prospect in STEM careers.

2. Literature Review

2.1 STEM and Non-STEM Stream

In tandem with the STEM action plan, STEM has been given great emphasis in the Malaysian education system by reforming its curriculum to prepare students for future STEM academic pursuits or careers. According to Mokhtar (2019), a "streamless" curriculum would replace the traditional science and arts stream through the latest national curriculum. Nevertheless, students are still divided into STEM and Non-STEM stream to establish a STEM-oriented education system so that all students in the upper secondary schools (equivalent to high school) can learn at least one STEM subject based on their preferences (Curriculum Development Division, 2016). In other words, STEM stream students learn more advanced STEM knowledge and skills, while non-STEM students are also given the option to register one STEM subject upon their enrolment at the upper secondary school (Ali et al., 2021; Curriculum Development Division, 2016). STEM subjects introduced to the students were designed based on industrial demand to train adequate talents for STEM workforce (Mokhtar, 2019). However, the implementation of the STEM-oriented streaming system is still at its infancy that it has been officially implemented nationwide since 2020 (Mokhtar, 2019), thus it is not clear if the new streaming system would result in differences between the two streams of students.

Xu (2013) is one of the very rare past research that examined the difference between STEM and non-STEM students. It was reported by Xu (2013) that there were differences in the factors that affected STEM and non-STEM graduates' career. Specifically, students were split into different streams, thus it can be expected that students from each respective field of study would choose a career that is align with their academic training (Xu, 2013). There are very limited career studies that examined students' differences in across streams of study, particularly in STEM and non-STEM. This raised the researchers' curiosity whether students from both the streams share similar perception towards work environment and prospect for STEM careers or not. For this reason, the present study investigated whether perceived work environment and career prospect in STEM would vary between STEM and non-STEM stream students.

2.2 Perceived Work Environment

In the review of literature, it was found that work environment could refer to various aspects such as safety to employees, job security, good relations with colleagues, and working hours (Raziq & Maulabakhsh, 2015). According to Raziq and Maulabakhsh (2015), work environment is a combination of work and context which includes various characteristics of an occupation is carried out and completed. Among the characteristics included in their research were working hours, job security and safety, relationships with colleagues, esteem needs, and the role of top management (Raziq & Maulabakhsh, 2015).

On the other hand, Crilly et al. (2017) suggested that work environment included four attributes, namely self-realisation, nervousness, workload, and conflict. In Sugahara and Boland's (2009) study that compared accounting and non-accounting students, they reported that there was a slight difference in terms of how the work environment influenced their career choice. Based on their research, sufficiency of social life, length of work hours and physical work conditions were the main aspects of work environment. As students are yet to experience an actual work experience, this study selected the aspects that would better relate to students' expectations to examine their perceived work environment. Therefore, perceived work environment in this study refers to a secondary school student's expected combination of work contexts in STEM careers that include sufficiency of social life, relationships with colleagues, length of work hours, application of knowledge and skills, and physical work conditions. Given the discussions, the following hypothesis was proposed:

H1: There is a significant difference between STEM and non-STEM students' perceptions towards work environment in STEM.

2.3 Career Prospect

Career prospect is defined as an individual's perception of the promotion opportunities offered at workplace, advantages working in the industry, and academic qualification as a worthwhile investment in career development (Wan et al., 2014). Based on Liaw et al. (2017), career opportunity, career stability, and good income are the three important attributes of career prospect. Hence, when a person has a positive prospect for a career, it is expected that the career would ensure a desirable income and living standard, provide opportunities to work overseas and career advancement, as well as to attain higher academic qualification (Liaw et al., 2017).

Wan et al. (2014) reported that career prospect is a crucial factor that influences students' commitment in their targeted career. Quansah et al. (2020) explained that when an individual places more optimistic prospect on the targeted career, the individual is more likely to choose a career in the targeted field. In Sugahara and Boland (2009), it was found that there was a slight difference in accounting and non-accounting students' career prospect for accounting career choice. Although these students perceived career prospect as an important key that led to their career decision in accounting, career prospect was only significant for non-accounting students when the career is expected to generate good long-term income. In this study, career prospect refers to a secondary school student's perception towards a STEM career in terms of aspects such as career stability, promotion opportunities, academic advancement, and social prestige. In view of the above discussions, the following hypothesis was formulated:

H2: There is a significant difference between STEM and non-STEM students' perceptions towards career prospect in STEM.

3. Research Methods

This study employed a survey design using a bilingual questionnaire to test the proposed hypotheses. It was a bilingual questionnaire in English and Malay language (the national language of Malaysia). The first section of the questionnaire included the respondents' demographic information such as the location of schools and stream of study. Perceived work environment and career prospect were measured in the subsequent section with six and eight items, respectively. Each item in the constructs were measured using a five-point Likert scale ranging from 1 = Strongly Disagree to 5 = Strongly Agree. The Cronbach's Alpha values for perceived work environment was 0.87 and career prospect was 0.91, which were above the recommended value of 0.70 (Pallant, 2013). In this study, the scales of both constructs consisted of existing items adapted from the existing literature (Crilly et al., 2017; Liaw et al., 2017; Sugahara & Boland, 2009; Wan et al., 2014), as well as self-developed items based on the need of the study.

Prior to data collection, the researchers had been granted the permissions from the Malaysian Ministry of Education, state offices of education, and the researchers' affiliation Scientific and Ethical Review Committee to conduct research. The purpose of the study was explained in the consent forms for the participants and their parents. Parental consent was sought because the respondents were 16-year-old adolescents who were considered under parents' care in Malaysia. Consequently, the respondents of this study were 157 students from the three states (Kelantan, Pahang, and Terengganu) in the East Coast of Malaysia. There were 60 students from Kelantan, 50 of them from Pahang, and 47 of them from Terengganu. More than half of the sample of this study were female students (64.3%, n = 101), whereas the remaining 35.7% (n = 56) of them were male. Among the respondents, 60 students indicated that they were from STEM stream, while the other 97 were non-STEM stream students. All the respondents in this study were 16-year-old, Form Four secondary school students (equivalent to Grade 10).

4. Results

4.1 Independent-Samples t-Tests

The collected data were analysed using independent-samples t-tests with IBM SPSS 23. Skewness and kurtosis were assessed, and the values were within the accepted range of ±1.00 (Hair et al., 2017). According to Pallant (2013), an independent-samples t-test is used to compare the mean scores of two distinct groups of continuous variables. Besides, Pallant (2013) noted that the effect size offers an indication of the magnitude of the differences between the compared groups. In this study, eta squared was used to represent the proportion of variance in the dependent variables that can be explained by the independent variable. Eta squared was calculated using the formula $t^2/ (t^2 + df)$. Eta squared < .06 indicates small effect, eta squared < .14 indicates moderate effect, while eta squared ≥ .14 indicates large effect (Pallant, 2013).

Align with the H1, an independent-samples t-test was conducted to compare the perceived work environment scores for STEM and non-STEM students. Based on the results as shown in Table 1, the significance level of Levene's test for equality of variances was larger than .05. Hence, it was assumed that there were equal variances between STEM and non-STEM stream (Pallant, 2013). The results generated from the analysis showed that there was a significant difference in scores for STEM (M = 3.89, SD = 0.74) and non-STEM (M = 3.49, SD = 0.75) students, with *t* (155) = 3.30, *p* < .05. The magnitude of the differences in the means (mean difference = .41, 95% confidence interval: .16 to .65) was very small (eta squared = .006), suggesting that only 0.6% of the variance in perceived work environment was explained by stream of study.

Table 1. *Independent Samples Test for Perceived Work environment*

Levene's Test for Equality of Variances		t-test for Equality of Means						
				Sig.	Mean	Std. Error	5% Confidence Interval of the Difference	
F	Sig.	t	df	(2-tailed)	Difference	Difference	Lower	Upper
.003	.955	3.297	155	.001	.40475	.12277	.16224	.64726

Besides, an independent-samples t-test was also carried out to test H2, whether there was any significant difference between STEM and non-STEM students' perceived career prospect. Table 2 shows that the significance level of Levene's test for equality of variances was .629 which was above .05, indicating there were equal variances between the two groups (STEM and non-STEM stream). From the results, it was shown that there was a statistically significant difference in STEM (M = 3.97, SD = .73) and non-STEM (M = 3.69, SD = .76) students' perceived career prospect in terms of their stream of study, with *t* (155) = 2.30, *p* < .05. The magnitude of the differences in the means had small effect (mean difference = .28, 95% confidence interval: .04 to .52, eta squared = .03). This means that 3.0% of the variance in perceived career prospect was explained students' stream of study.

Table 2. *Independent Samples Test for Perceived Career Prospect*

Levene's Test for Equality of Variances		t-test for Equality of Means						
				Sig.	Mean	Std. Error	95% Confidence Interval of the Difference	
F	Sig.	t	df	(2-tailed)	Difference	Difference	Lower	Upper
.235	.629	2.304	155	.023	.28269	.12271	.04030	.52508

5. Discussion and Conclusion

5.1 Discussion

In Malaysia, the reformation of the education system was enacted to reinstate the importance of STEM components in its curriculum by offering STEM subjects to all students at the upper secondary level. The goal of this initiative is to better prepare students from all streams of study for careers in STEM. However, the findings indicated that despite having made “STEM for all” in the latest curriculum, students from both streams of study perceive the work environment and career prospect in STEM differently. Though the effects of the results were small, this finding would offer insights on research implications and future studies.

Based on the findings, this study found a need to view students’ career perception towards work environment and career prospect in career decision separately despite the “streamless” system. Xu (2013) mentioned that since students are separated into STEM and non-STEM majors, the system per se has isolated the given exposure and training that qualify a student for a STEM career. The unequal exposure and training that STEM and non-STEM students obtain also add to the difficulties for non-STEM candidates to seek a career in STEM fields when the competitors are well-equipped with the qualifications needed for a STEM career. As such, students are more likely to choose careers that are related to their major (Xu, 2013).

Although the curriculum was reformed to emphasise STEM subjects, and has made STEM available to non-STEM students, there is still a gap in terms of the amount of STEM knowledge they learn at schools that equip them for STEM careers. STEM stream students are prepared to pursue STEM professionals, whereas the non-STEM students are expected to pursue careers in humanities and arts, with the alternate option to seek jobs as STEM associates if they choose to learn STEM subject in school. As such, the amount of STEM components and exposure in both streams of studies has naturally segregated non-STEM students from STEM. This could be a reason that has led to the difference in students’ perception towards work environment and career prospect in STEM in which STEM students are given more opportunities to learn about STEM than non-STEM students.

The STEM students were given sufficient exposure and training about STEM careers, they had more specific expectations of the work contexts in STEM careers such as the length of work hours, and physical work conditions. On the other hand, STEM careers were regarded as alternate option for non-STEM students where they were not provided much exposure as STEM students. Hence, non-STEM students’ perception towards the work environment and work contexts in STEM careers might not be as evident as the STEM students. Likewise, as students were split into different streams, STEM students were expected to join the workforce as STEM professionals while non-STEM students were more likely to be STEM associates if they pursue STEM careers. This could have led to the differences in terms of STEM and non-STEM students’ expectations towards STEM careers such as income, career advancement, and opportunities. Due to the differences in knowledge and skills required for the two categories of STEM jobs, students from STEM and non-STEM stream tend to have different prospects for STEM careers.

5.2 Limitations, Recommendations & Conclusion

The findings of the present study would contribute to understanding of the new streaming system in the Malaysian national curriculum. The arts and science streaming system were abolished to consolidate its education system and make STEM available for all students in the upper secondary level. However, the classification of students into STEM and non-STEM is still a form of academic streaming that discerns the learning contents and trainings for students. Considering the small effect of the results, this study could still be a meaningful reference to the policy makers to emphasize STEM opportunities for non-STEM students, so that they are not isolated from STEM despite having the chance to enter the STEM industries (e.g.: STEM associates). Additionally, the new streaming system is still at the early stage of its implementation. This study offers insights about the most updated STEM scenario to the direct consumers of the education system, STEM and non-STEM students.

There are a few limitations in this study that should not be overlooked. The results generated from this study can only represent students from the East Coast of Malaysia. This limitation has restricted the generalizability of the research finding, thus future studies can expand the research scope by including students from other regions of Malaysia, especially those from the central region. According to the Ministry of Education (2013), the central region of Malaysia is the education hub where access and resources are generally given the utmost priority. On the contrary, students from other regions such as the East Coast might have less opportunities and access about work environment and career prospect in STEM due to access restrictions. Pertaining to this, it is also suggested that future research could use other variables such as career interest and career choice intention to test the differences between STEM and non-STEM stream students.

Acknowledgements

We are grateful to Universiti Tunku Abdul Rahman Research Fund (UTARRF) for funding and supporting this research.

References

Ali, G., Jaaffar, A.R. and Ali, J. (2021). STEM education in Malaysia: Fulfilling SMEs' expectation. In Sergi, B. S. & Jaaffar, A. R. (Eds.), *Modeling Economic Growth in contemporary Malaysia* (*entrepreneurship and global economic growth*) (pp. 43-57). Emerald Publishing Limited, https://doi.org/10.1108/978-1-80043-806-420211005

Crilly, J., Greenslade, J., Lincoln, C., Timms, J., & Fisher, A. (2017). Measuring the impact of the work environment on emergency department nurses: A cross-sectional pilot study. *International emergency nursing, 31*, 9-14. https://doi.org/10.1016/j.ienj.2016.04.005

Curriculum Development Division. (2016). *Buku penerangan kurikulum standard sekolah menengah* (KSSM). Ministry of Education Malaysia. http://bpk.moe.gov.my/index.php/terbitan-bpk/buku-penerangan-kssr-kssm?download=1720:buku-penerangan-kssm

Hair, J. F., Hult, G. T. M., Ringle, C. M., and Sarstedt, M. (2017). A *primer on partial least squares structural equation modeling* (2nd ed.). Sage.

Liaw, S. Y., Wu, L. T., Chow, Y. L., Lim, S., & Tan, K. K. (2017). Career choice and perceptions of nursing among healthcare students in higher educational institutions. *Nurse education today*, *52*, 66-72. http://dx.doi.org/10.1016/j.nedt.2017.02.008

Ministry of Education. (2013). *Malaysia education blueprint 2013-2025*. Putrajaya: Ministry of Education, Malaysia. Retrieved from https://www.moe.gov.my/muat-turun/penerbitan-dan-jurnal/dasar/1207-malaysia-education-blueprint-2013-2025/file

Mokhtar, H. S. (2019, November 20). Streamless upper secondary next year: What it actually means. *New Straits Times.* https://www.nst.com.my/education/2019/11/540506/streamless-upper-secondary-next-year-what-it-actually-means

Pallant, J. (2013*). SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (5th ed.). Open University Press, McGraw-Hill Education.

Quansah, F., Ankoma-Sey, V. R., & Dankyi, L. A. (2020). Determinants of Female Students' Choice of STEM Programmes in Tertiary Education: Evidence from Senior High Schools in Ghana. *American Journal of Education and Learning*, *5*(1), 50-61. http://dx.doi.org/10.20448/804.5.1.50.61

Raziq, A., & Maulabakhsh, R. (2015). Impact of working environment on job satisfaction. *Procedia Economics and Finance*, *23*, 717-725. http://dx.doi.org/10.1016/S2212-5671(15)00524-9

Shahali, E. H. M., Ismail, I., & Halim, L. (2017). STEM education in Malaysia: Policy, trajectories and initiatives. *Asian Research Policy, 8*(2), 122–133. http://www.arpjournal.org/usr/browse/list_issues_detail.do?seq=27

Sugahara, S., & Boland, G. (2009). The accounting profession as a career choice for tertiary business students in Japan-A factor analysis. *Accounting Education: an international journal*, *18*(3), 255-272. http://dx.doi.org/10.1080/09639280701820035

Wan, Y. K. P., Wong, I. A., & Kong, W. H. (2014). Student career prospect and industry commitment: The roles of industry attitude, perceived social status, and salary expectations. *Tourism Management*, *40*, 1-14. https://doi.org/10.1016/j.tourman.2013.05.004

Xu, Y. J. (2013). Career outcomes of STEM and non-STEM college graduates: Persistence in majored-field and influential factors in career choices. *Research in Higher Education*, *54*(3), 349-382. http://dx.doi.org/10.1007/s11162-012-9275-2

Zhang, L.-f., Fu, M., & Li, D. T. (2020). Hong Kong academics' perceived work environment and job dissatisfaction: The mediating role of academic self-efficacy. *Journal of Educational Psychology, 112*(7), 1431–1443. https://doi.org/10.1037/edu0000437

Fostering Conceptual Change in Software Design

Lakshmi T G* & Sridhar IYER
IDP in Educational Technology, Indian Institute of Technology Bombay, India
*tglakshmi@iitb.ac.in

Abstract: Novices and experts exhibit differences in understanding and creating software conceptual design. Experts are known to build integrated software solutions that fulfill the requirements of real-world design problems. Novices have specific difficulties such as fixation and lack of integration while creating software conceptual design. Software design teaching-learning approaches have been directed towards software methodologies, processes, and tools. However, teaching-learning interventions for specific novice difficulties and disciplinary practices of software conceptual design are still not available. We created a function-behaviour-structure (FBS) based learning environment, 'think & link' to alleviate the novice difficulties and disciplinary practices of integrated software solution creation. The aim of the study in this paper is to examine the change in novices' outcome and understanding of software conceptual design after having completed all activities in 'think & link'. The study was conducted with final year undergraduate computer and information technology students (n=20) from an engineering college in Mumbai, India. There is no gestalt shift in the pre-post solutions to design problems. However, learners' conceptions of software conceptual design indicate conceptual change. The learners' refined their understanding and developed perspectives about software conceptual design.

Keywords: Conceptual change, software engineering disciplinary practices, function-behaviour-structure design framework, technology enhanced learning environment

1. Introduction

Conceptual design activity is described as a process in which the functional requirements of the design problem are extracted and transformed into descriptions of solution concepts (Chakrabarti & Bligh, 2001). The conceptual phase of design is significant, as designers tend to develop numerous early ideas and solutions in this phase. Conceptual design is inherently hard and needs to be supported (Chakrabarti & Bligh, 2001). The characteristics of software such as intangibility and dynamicity add to the complexity of software conceptual design (Petre et al, 2010).

In the context of software conceptual design (SCD) it is a standard practice to create various representations of unified modeling language (UML) to represent the solution design. However most of the designs created using UML describe system in different notations from different points of view and at different levels of abstraction. A SCD is described by integrating the various UML representations. There is a lack of a single representation to represent all views (Niepostyn & Bluemke, 2012). In the formal curricula of computer engineering and information technology, students learn about syntax, semantics and processes to create the formal (UML) representations. However when students encounter open-ended real world problems they are unable to utilize the formal representations or create meaningful SCD (Eckerdal et al, 2006). Novices are unable to utilize multiple integrated representations in UML for a given design problem (Eckerdal et al., 2006; Lakshmi & Iyer, 2018).

To alleviate these novice difficulties while creating SCD we designed and developed a function-behaviour-structure (FBS) design framework (Gero & Kannengiesser, 2014) based learning environment - 'think & link'. In 'think & link' FBS framework manifests as a manipulable graph. This paper, describes the pedagogical design and features of 'think & link'. The evaluation of the effect of 'think and link' is examined using the conceptual change lens. We examine the learners' pre-port SCD designs and their understanding of SCD after they have completed activities in 'think & link'.

2. Background

Conceptual change research begins with addressing learners' understanding of a given topic and the change in understanding as a result of instruction (von Aufschnaiter & Rogge, 2015). In the context of computer science educators it assists in understanding novices' conception and creating specific instruction to address alternate conceptions (Qian & Lehman, 2017).

Prior studies on novice difficulties in SCD (Eckerdal et al, 2006) (Thomas et al, 2014) (Chren et al, 2019) indicate that novices (i) only rewrite problem statements during design phase, (ii) are unable to utilize formal representations of UML to model SCD and (iii) are unable to utilize multiple UML diagrams for integrated view of solution. Our studies confirmed these findings (Lakshmi & Iyer, 2018). Our studies have further informed us that novices face the difficulties of fixation and lack of integration during the SCD task. For example, given the design problem of mood based music player our study with novices' (Lakshmi & Iyer, 2018) found that they mostly focused on mood detection. The fixation was towards a specific function, structure or behaviour. In contrast expert software designers create comprehensive and cohesive SCD by understanding the problem and creating integrated solution view (Ball et al, 2010). Creating integrated solution designs involves– i) utilization of multiple UML representations that are linked and ii) addressing both problem and solution aspects of the given design problem.

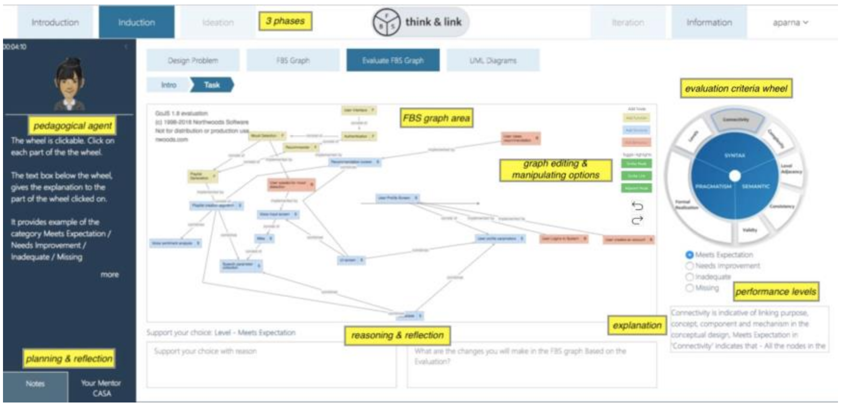
The FBS framework (Gero & Kannengiesser, 2014) models designing in terms of the design elements: function (F), behaviour (B), structure (S), expected behaviour (Be) and behaviour derived from structure (Bs). Functions, describe what the design is for; behaviours, describe what it does; and structures, describe what it is. Along with FBS elements the framework has 2 sets of behaviours, expected behaviour (Be) and behaviour derived from structure (Bs). One may distinguish two main fields of application, '*as a theoretical vehicle for understanding design, and as a conceptual basis for computerized tools intended to support practicing designers* '(Galle, 2009). The function-behaviour-structure (FBS) design framework (Gero & Kannengiesser, 2014) can be considered an appropriate framework to alleviate novices' difficulties of integrating representations in SCD.

3. Design of 'think & link

'think & link' is a web-based, self-paced, FBS framework based learning environment for teaching-learning of SCD. The FBS framework manifests as a FBS graph in the learning environment. The FBS graph is a representation through which learners can symbolize function, structure, behaviour and establish the relationship between them. The FBS graph pedagogy includes creation and manipulation of a representation. 'think &link' consists of scaffolds for learners to create, modify and evaluate a FBS graph for design problems. There are three phases in 'think & link' and the following features:

- FBS graph manipulator and editor – This feature is present throughout the three phases. However in the first phase alone the editor options are not provided to the learners. The graph manipulator displays the FBS graph for the problem with color-coded nodes. The clickable options on the right panel help the learner to display similar nodes, links and adjacent nodes (see figure. 1). In the edit mode the right panel extends clickable options to add function, structure and behaviour nodes. Dragging the cursor from the source node and placing it on destination node creates the link. The link can be annotated with tags - 'implemented by, consists of, and combines' by right clicking on the link. Using the activity, manipulator and clickable options in phase 1 learners build FBS conceptual model. This conceptual model helps the learners link F/B/S together. By editing FBS graphs learners build strategies to create and establish links between F/B/S for a given design problem.
- FBS graph evaluator – This feature is present in the phase 2 and 3. It aids in the self-evaluation of FBS graph based on criterion of syntactic, semantic and pragmatic categories. The criteria of conceptual model were adapted from Lindland et al. (1994). The categories include properties like connectivity, complexity, consistency, validity, consistency, levels and formal realization. All these parameters are adapted and presented in the context of the FBS graph. The evaluation categories are presented as a clickable wheel (see figure. 1). Clicking on an evaluation criteria the performance levels (meets expectation, needs improvement, inadequate, missing) are presented as radio buttons.

The explanation of the criteria and the respective selected performance level is presented to the learner. The learner has to select the performance level after evaluating the FBS graph. The learner needs to support the performance level choice with reason and state the corresponding changes in FBS graph that the learner would make. The learners' self evaluate the categories of SCD in the context of FBS graph. Learners are also required to provide reasoning for the evaluation and reflect on the changes in FBS graph.



The wheel is clickable. Click on each part of the the wheel.

The text box below the wheel, gives the explanation to the part of the wheel clicked on.

It provides example of the category Meets Expectation / Needs Improvement / Inadequate / Missing

GoJS 1.8 evaluation
(c) 1998-2018 Northwoods Software
Not for distribution or production use
nwoods.com

● Meets Expectation
○ Needs Improvement
○ Inadequate
○ Missing

Connectivity is indicative of linking purpose, concept, component and mechanism in the conceptual design. Meets Expectation in 'Connectivity' indicates that - All the nodes in the

Support your choice: Level - Meets Expectation

Support your choice with reason

What are the changes you will make in the FBS graph Based on the Evaluation?

Figure 1. Features of 'Think & Link'.

- Pedagogical agent – On the left side of 'think & link' (see figure. 1) a vertical column is dedicated to the pedagogical agent CASA (conceptual design assistant). CASA is present all through the phases. CASA provides procedural prompts related to the task that the learner is currently performing. The learner's task progress is monitored and CASA provides appropriate scaffolds to complete the task at the desired level. CASA also provides cognitive prompts, which aid the learner in creating the FBS design elements and linking them.
- Resources for Information - At each and every step of tasks in 'think &link' there are videos, which contain task specific knowledge required to complete the task (see figure. 1). Additionally there is also 'Information' page in the learning environment that includes a collection of videos about the context of the learning environment like SCD, FBS framework etc.
- Planning questions – It is important for the learner to reflect, evaluate and monitor their process during design. By doing this learners will be able to imbibe the process of SCD along with the strategies. As mentioned earlier the learners are taken through the three planes of cognition – doing, evaluation and synthesis. In the planning activity the learners are required to reflect on the task ahead and plan (see figure. 1). Example questions that they encounter are – *What will you do in this phase? How will this task be useful for creating software conceptual designs?*

4. Research Method

The broad research question guiding this study is - What is the nature of conceptual change in software conceptual design after learners' initial interaction with the learning environment 'think & link'?

4.1 Research Design, Participants and Study Procedure

The study was conducted as a hands-on one-day workshop in an urban private engineering institute. The participants for the workshop were selected via purposive sampling. Only participants in their final year of computer engineering (CS) and information technology (IT) were considered for the workshop. 20 students registered for the workshop (CS=15, IT=5: male=16, female=4). The study participants are representative of Indian urban engineering students.

In the workshop registration form along with personal details, participants answered an open-ended questionnaire aimed to capture their prior conception of SCD. First author of the paper along with a colleague were present during the workshop. Participants solved a pre-test at start. They individually created a SCD on pen and paper for the design problem - 'Design a mood based automatic music player'. They were free to use the internet for this task. After completing this task, participants utilized the individual desktop to access ‘think & link’. ‘think & link’ has three phases which the participants completed in ~4.5 hours. After completing all activities in ‘think &link’, participants for an hour solved another SCD for – ‘Design a finger print ATM system’. The pre and post SCD problems can be considered equivalent, as they are similar in terms of complexity and time taken to solve. After completing the post-test, participants were asked to respond to questions about understanding of term software conceptual design and a focus group semi structured interview was conducted. Participants spent 5 hours in the workshop.

4.2 Data Sources and Analysis

The online registration form for the workshop contained prior conception open-ended question to capture participants’ understanding of the term software conceptual design. The question that participants were asked was, “What is your understanding of 'software conceptual design'?” The sketches and notes for the pre and post paper based activities were collated as artifacts of the design activity. At the end participants’ conceptions about software conceptual design, usability and usefulness of ‘think & link’ was also captured as responses to online questionnaires. Table 1 below maps the measure, data source and the analysis technique.

Table 1. *Mapping of Measure, Data Source and Analysis Technique*

Measure	Data Source	Analysis
Pre-post change in SCD	Artifacts	Category evaluation [(Thomas et al., 2014)
Pre and post change in understanding of SCD	Pre-post open ended responses	Inductive thematic analysis (Clarke & Braun, 2016)

The categories of software conceptual design by Thomas et al. (2014) were utilized to analyze the pre-post design artifacts. These categories were utilized to classify the pre and post SCD of participants. Once the participants’ artifacts were classified, there emerged a need to explicate the participants’ transition across the categories in pre and post-test. The tool interactive stratified attribute tracking, iSAT (Majumdar & Iyer, 2016) was used to analyze the pre-post categories transition.The visualization from the tool is shown in figure. 5 and elaborated in the next section.

To analyze the open-ended responses to the questionnaire the guidelines provided by Clarke and Braun (2016) were followed. The unit of analysis was a sentence. The codes represented the relevant qualities, activities and outcome of SCD. The comparison between the pre-post responses is captured in sub-section 5.1 and 5.2 in results section.

5. Results

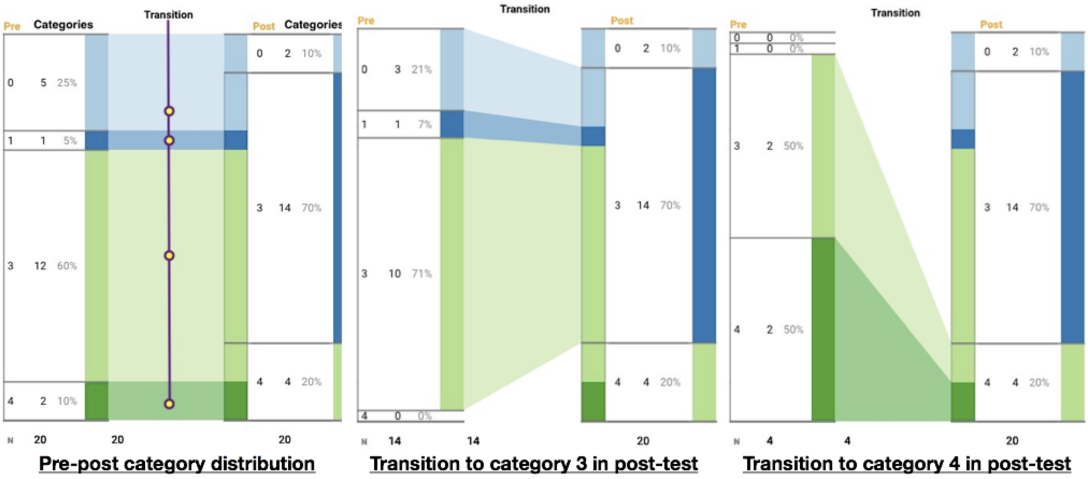
The goal of this paper is to explore the nature of conceptual change in SCD after learners completed activities in ‘think &link’. To analyze the transitions across the pre-post categories, we used the tool iSAT. The observations from analyzing the pre –post transitions (figure 2) are:

- none of the participant slid down to lower category in post test
- majority of the participants in both pre and post fell in the category 3 as they utilized the representation of flowchart
- 3 participants’ in the post intervention moved to the category 3 from category 0 (figure 4). The majority of the dynamic representations utilized by participants were flow chart, so that could explain the shift. A participant moved to category 3 from category 1. This movement could indicate that participants are able to create dynamic representations than the static formal representations.
- 2 participants’ in the post intervention moved to the category 4 from category 3 (figure 4). During the intervention the understanding that software conceptual design comprises of multiple artifacts depicting the functional, behavioral and structural view could be the reason for the shift.

There was no ‘gestalt shift’ in the participants’ artifacts, i.e. all the participants in post intervention did not move to category 4. However the slight shift as evident from drop in category 0, rise in category 3 and 4. Additionally to investigate the participant understanding of SCD the open-ended response thematic analysis was done. Two major themes of refinement of understand and perspective shift emerged. Under each subsection, the shift in understanding is discussed with example responses, their corresponding code and theme.

5.1 Refinement of Understanding about SCD

The participants' developed the understanding about SCD being a *'combination of all UML diagrams'* (code: drawings, theme: outcome) rather than just thinking about *'conceptual & schematic drawings'* (code: drawings, theme: outcome). There is a refinement in the outcome of SCD as an integrated solution from participants' responses such as- (i) *doing a conceptual design going deep into what actually the problem is and form connections to various solution parts actually*, (ii) *what will be the back end, what will be the front end, how will front end access back end*. The other observed refinement in participants' understanding about SCD is that *'.... need to understand intricacies for implementing minor details'* (code: details of solutions, theme: activities) whereas in pre response we see the response as *'creating design modules'* (code: module creation, theme: activities). Earlier participants' understanding about activities in SCD was 'documentation' whereas now they refine their understanding as SCD involving design and creation ('*in software engineering we didn't design anything, it was just documentation'*).



Pre-post category distribution **Transition to category 3 in post-test** **Transition to category 4 in post-test**

Figure 2. Transition of Pre-Post Artifact Category Classifications.

5.2 Perspective Shift

The participants' developed alternative views about SCD, which will be useful during solving design problems. The first shift in perspective is about SCD being a systematic approach – '*it is a systematic way instead of just throwing things on paper'* (systematic approach- activities). Pre-intervention participants' design for 'customer requirements in modules'. After intervention participants' acquire the perspective of designing for understanding of other '*designers as well as programmer or developer'*. The participants' also develop the perspective about the cognitive process involved in the activity – *'we need to first imagine how will end user use it, then create use cases, identify components'*. The participants' view SCD as a stage '*before coding'* where they are required to *'mention all steps so that it is as close to the real software',* whereas, in pre intervention response they view it as a '*phase extracting problem chara*cteristics'.

6. Discussion and Conclusion

The driving research question in this paper is to examine the nature of conceptual change in learners understanding of SCD. The pre-post open-ended responses indicate that learners' have undergone a shift in their understanding of SCD. The open-ended responses collated before the commencement of the workshop capture the conception participants had before the exposure to 'think & link'. Although the coding themes in pre-post almost remain the same, the responses themselves were contrasting in nature. This indicates a conceptual integration (Vosniadu, 2013), in which the practices and understanding from the activities in 'think & link' are combined with participants' earlier conceptions. Although the pre-post change in categories of design solution observed at the end of the workshop was not significant (figure 2). The gradual change in the participants' understanding about the outcomes and activities in SCD is indicative of conceptual change (Nussbaum, 1989). The design features of 'think &link' could be utilized by teachers/researchers who want to develop conceptual change in novices' understanding of SCD. The results from this study have implication for the teaching learning of software conceptual design. The design features of 'think &link' could be utilized by teachers/researchers who want to develop conceptual change in novices' understanding of scd. Some of the results that can be utilized for teaching learning of SCD are: (i) explicit linking of different aspects of scd, (ii) recognizing importance and deliberate practice of SCD, (iii) scaffolds for cognitive processes. We have identified two major limitations of the study. First, the sample size of the study is small (n=20) for generalizability. However the goal of this study is to explore the nature of conceptual change in the context of software conceptual design after learner interactions with the mediating tool 'think and link'. Second, reconfirmation of conceptual change by participants after thematic analysis was not done.

Future studies have been planned to closely examine the changes. Additionally, longitudinal studies have been planned by providing access to the learners to interact with 'think & link' for longer periods of time and collect data pertaining to processes, practices and discourse.

References

Ball, L. J., Onarheim, B., & Christensen, B. T. (2010). Design requirements, epistemic uncertainty and solution development strategies in software design. *Design Studies*, 31(6), 567-589.

Chakrabarti, A., & Bligh, T. P. (2001). A scheme for functional reasoning in conceptual design. *Design Studies*, *22*(6), 493-517.

Chren, S., Buhnova, B., Macak, M., Daubner, L., & Rossi, B. (2019, May). Mistakes in UML diagrams: analysis of student projects in a software engineering course. In *Proceedings of the 41st International Conference on Software Engineering: Software Engineering Education and Training* (pp. 100-109). IEEE Press

Clarke, V. & Braun, V., 2016. Thematic analysis. *The Journal of Positive Psychology*, 12(3), pp.297-298

Eckerdal, A., McCartney, R., Moström, J. E., Ratcliffe, M., & Zander, C. (2006). Can graduating students design software systems?. In *SIGCSE'06* (pp. 403-407). ACM.

Galle, P. (2009). The ontology of Gero's FBS model of designing. *Design Studies*, *30*(4), 321-339.

Gero, J. S., & Kannengiesser, U. (2014). The function-behaviour-structure ontology of design. In *An anthology of theories and models of design* (pp. 263-283). Springer, London.

Lakshmi, T. & Iyer, S. (2018). Exploring Novice Approach to Conceptual Design of Software. *In Kay, J. and Luckin, R. (Eds.) Rethinking Learning in the Digital Age: Making the Learning Sciences Count, 13th International Conference of the Learning Sciences (ICLS) 2018*, Volume 3. London, UK: International Society of the Learning Sciences

Lindland, O. I., Sindre, G., & Solvberg, A. (1994). Understanding quality in conceptual modeling. *IEEE software*, *11*(2), 42-49.

Majumdar, R., & Iyer, S. (2016). iSAT: a visual learning analytics tool for instructors. *Research and practice in technology enhanced learning*, *11*(1), 16.

Niepostyn, S. J., & Bluemke, I. (2012, June). The Function-Behaviour-Structure Diagram for Modelling Workflow of Information Systems. In *International Conference on Advanced Information Systems Engineering* (pp. 425-439). Springer, Berlin, Heidelberg.

Nussbaum, J. (1989). Classroom conceptual change: philosophical perspectives. *International Journal of Science Education*, 11(5), 530-540.

Petre, M., van der Hoek, A. and Baker, A. (2010). Editorial. *Design Studies*, 31(6), pp.533-544.

Qian, Y., & Lehman, J. (2017). Students' misconceptions and other difficulties in introductory programming: A literature review. *ACM Transactions on Computing Education (TOCE)*, *18*(1), 1.

Thomas, L., Eckerdal, A., McCartney, R., Moström, J. E., Sanders, K., & Zander, C. (2014, July). Graduating students' designs: through a phenomenographic lens. In *Proceedings of the tenth annual conference on International computing education research* (pp. 91-98). ACM.

Von Aufschnaiter, C., & Rogge, C. (2015). Conceptual change in learning. *Encyclopedia of science education*, 209-218.

Vosniadou S (ed) (2013) International handbook of research on conceptual change, 2nd edn. Routledge, New York/London

The Effectiveness of Collaborative Concept Map Recomposition and Discussion with Kit-Build Concept Map in Online Learning

Aryo PINANDITO[a,b*], Didik Dwi PRASETYA[a,c], Nawras KHUDHUR[a], Yusuke HAYASHI[a] & Tsukasa HIRASHIMA[a]
[a]*Information Engineering, Graduate School of Engineering, Hiroshima University, Japan*
[b]*Information System Department, Faculty of Computer Science, Universitas Brawijaya, Indonesia*
[c]*Information Technology Department, Faculty of Engineering, Universitas Negeri Malang, Indonesia*
*aryo@ub.ac.id

Abstract: Collaborative learning with concept maps has been recognized as a learning tool that positively impacts learning. Recently, transforming the learning environment into online settings becomes a priority to face the shift towards distance learning in the current pandemic. Learning with Kit-Build concept map is also no exception. Finding the factors that influenced student's comprehension in online collaborative learning with Kit-Build concept map could help teachers optimize their teaching strategy and aid students to learn in a better way; thus, improve the desired learning outcomes. In the context of online learning with concept maps, this study aims to investigate whether student's comprehension is affected by the direct influence of concept mapping strategy or the activeness of the discussion during collaboration. The results suggested that Kit-Build concept map encouraged better discussions, influenced and improved students' comprehension better than the traditional open-ended concept mapping.

Keywords: Collaborative learning, concept map, Kit-Build, learning effect, online

1. Introduction

Concept maps can be used to visualize ideas, depict relationships between two or more concepts, and structure knowledge. Learning with concept maps allows teachers and students to construct their understanding of concepts and relationships logically and in a structured sense. Therefore, a concept map has been acknowledged as an alternative tool for teaching, learning, assessment (Hirashima et al., 2011; Hirashima et al., 2015), exploring knowledge in research (de Ries et al., 2021). Furthermore, concept maps can be presented in digital forms and provided in an online environment (Metcalf et al., 2018) to improve its applicability in distance learning.

Elaborating concept maps into collaborative learning may cultivate deeper learning (Chen et al., 2018) and enhance critical thinking skills (Tseng, 2020) in conflict resolutions. Kit-Build is one learning framework that uses a concept map recomposition strategy to help students understand learning materials better (Hirashima et al., 2015). Collaboratively recomposing concept maps with Kit-Build could promote more active discussions and encourage students to share their understanding better than the traditional open-ended concept mapping (Wunnasri et al., 2018).

Many factors could affect student understanding during collaborative learning. For example, active discussion and participation of students were found to affect student understanding (Dallimore et al., 2016; van Blankenstein et al., 2011). Finding the factors that influenced their learning and further improved the quality of said factors could help teachers revamp their teaching strategy and the quality of teaching materials. A study in online collaborative learning (Pinandito et al., 2021) suggested that Kit-Build concept map could encourage more active and engaging discussion among students. As a result, it helped students to understand and comprehend better than using an open-ended concept mapping approach. However, their activeness in the discussion during learning could also be the factor that influenced student understanding. This study investigated factors that could affect group comprehension during collaborative learning with Kit-Build concept map. To further guide this study, the following research questions were addressed:

1. Is using the Kit-Build concept map method improve group comprehension better than the open-end concept mapping method in online collaborative learning with concept maps?
2. Is the activeness of students in the discussion affect group comprehension? If so, what kind of talks in the discussion influenced their comprehension?

The result suggested that the Kit-Build concept map method could improve group comprehension better than the traditional open-end approach in an online collaborative learning environment. In addition to encouraging more active and engaging discussions, Kit-Build also encourages students to discuss the problem more, thus affecting their comprehension and memory retention positively.

2. Online Collaborative Learning with Kit-Build Concept Map

Kit-Build concept map (Kit-Build) is a learning method that uses concept map re-composition strategy in its learning activities. In learning with concept maps, students or teachers compose a concept map of a learning topic and help readers to understand the ideas quickly. Instead of composing a concept map from an empty workspace (scratch mapping), Kit-Build uses the re-composition style by providing a predefined set of concept map components—a kit—to recompose (Sugihara et al., 2012). This re-composition activity is called kit-building. In kit-building, students are guided to recompose their concept maps to a structure similar to their teacher's concept map in the form of feedback. Even though kit-building is more restrictive than scratch mapping, it is less cognitive-demanding (Tseng, 2020).

The kit of a Kit-Build concept map held the critical key of learning with it. It guides the students to focus on particular concepts and ideas of a learning topic and helps them comprehend the topic better. Kit-Build concept map can also be used as a formative assessment in learning (Yoshida et al., 2013, Pailai et al., 2017) by its feedback and automatic comparison mechanism (Hirashima et al., 2015). Extensions and variations of learning with Kit-Build concept map have been conducted in several studies. Extending Kit-Build with scratch mapping could help the students comprehend the learning material and compose their concept maps better (Prasetya et al., 2021). Support for concept map composition through a semi-automatic concept map authoring system was given to improve the concept mapping efficiency (Pinandito, Prasetya, Hayashi, & Hirashima, 2021b, 2021a).

In previous studies (Andoko et al., 2020; Pinandito et al., 2021), Kit-Build effectively supports learning English as a Foreign Language (EFL) reading comprehension. Using Kit-Build in collaborative learning also showed a positive learning effect while also encouraged the students to discuss more actively (Wunnasri et al., 2018, Pinandito et al., 2021). The system has been enhanced further (Pinandito, Prasetya, Az-zahra, et al., 2021) to support real-time collaboration and online use. Additionally, the system allowed the students to discuss with a unique node-related text-based communication interface, separating general discussions from concept-or-link discussions and help them manage and keep the discussion in control when discussing several topics or ideas at the same time. However, it is yet to be confirmed whether the factor affecting group comprehension is the concept mapping activity, their activeness in the discussion during collaboration, or both; thus, addressed in this study.

3. Methodology

This study used the term Scratch Mapping and Kit-Building to differentiate two concept map composition methods. Scratch Mapping represents concept mapping activity where concept map authors can freely compose concept maps from scratch. On the other side, Kit-Building represents concept map recomposition activity from a predefined set of concept map components—a Kit-Build concept map kit.

An experiment, as shown in Figure 1, was designed to answer the research questions. Before the actual collaboration activity, the participants were given a tutorial about concept maps and training on creating a good concept map from English reading with the Kit-Build system directly. Before the concept mapping, they read a 900-words of English text about “Wagyu, a Japanese Breed Cow” and take a 10-minute pre-test. After the collaboration, the students take a 10-minute post-test. During the 30-minute concept mapping, they have access to the reading text, but they neither could see nor talk with their partner to communicate directly. They can only talk with their partners using the communication channel provided within the system.

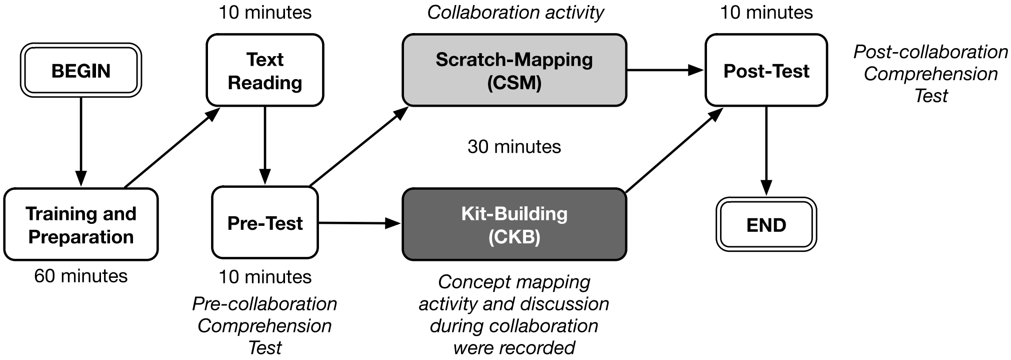


Figure 1. Experiment Flow of Collaborative Learning with Concept Maps.

The participants were 36 graduate students from Hiroshima University and four students from Miyazaki University, Japan. They were non-Japanese international students whose mother language is not English and used English as their primary language during their study in Japan. The students were forming groups of pairs (dyads) and divided into two groups, i.e., Scratch Mapping (CSM) and Kit-Building (CKB). To form the pairs, every student invited one friend at the same education level as their collaboration partner to eliminate language and communication problems during collaboration. According to the demographic questionnaire given, all participants have a minimum Test of English as a Foreign Language (TOEFL) Institutional Testing Program (ITP) equivalent score of 500. Therefore, they were assumed to have adequate English proficiency, especially in reading.

One Kit-Build concept map was made by an English teacher that consisted of 20 concepts and 19 propositions. The concept map represented the knowledge structure of the reading text, decomposed into a Kit-Build kit in a complete decomposition manner. The tests used the same set of questions that consisted of eight multiple-choice questions, and the questions were provided in random order.

4. Result and Discussion

4.1 Student Comprehension and Group Discussion

The pair group comprehension score was defined as the average individual test score of each pair group member. The pre-test and post-test scores were normalized to a maximum score of 10 and are shown in Table 1. The pre-test score data were analyzed with the Levene and Shapiro-Wilk tests to evaluate the homogeneity of variance and normality, respectively. The tests suggested that they were homogeneous (*p-value* = 0.08029 > 0.05) but fail to conform the normality (*p-value* = 0.02513 < 0.05). Therefore, non-parametric analysis methods were used in this study.

In addressing the first research question, whether the Kit-Building method improves group comprehension better than scratch mapping, the Mann-Whitney U tests were carried out to compare the students' pre-test and post-test scores. According to the Mann-Whitney U test *p-values* shown in Table 1 and Figure 2, all students have similar understanding levels before collaboration. However, the difference in both groups' understanding levels after the collaboration activity was statistically significant. The CKB group has better comprehension than the CSM group after the collaboration. Therefore, it can be said that the kit-building method could improve student comprehension better than scratch mapping.

Table 1. *Group Comprehension Test Score*

Test	Approach	n	Mean	Min	Max	Std. Dev.	Median	*p-value*	Sig.
Pre-Test	Kit-Building	10	4.25	3.75	5	0.493	4.38	0.7869	n.s.
Pre-Test	Scratch Mapping	10	4.29	2.14	5.71	1.12	4.64		
Post-Test	Kit-Building	10	7.62	6.88	9.38	0.768	7.5	0.0012	**
Post-Test	Scratch Mapping	10	6.14	5	7.86	0.768	6.07		

The system captured each message sent by the students and counted the message as one talk. Each talk was labeled with one category of the Advanced Interaction Analysis for Teams (act4teams) coding scheme (Kauffeld et al., 2018). The distribution of the talks is also shown in Figure 2. According to the classification, students of the Kit-Building group discuss more the problem because the students of the Kit-Building group have a kit to discuss since the beginning of the collaboration activity. On the contrary, students of the Scratch Mapping group talk about procedural matters more than discussing the topic (Pinandito et al., 2021).

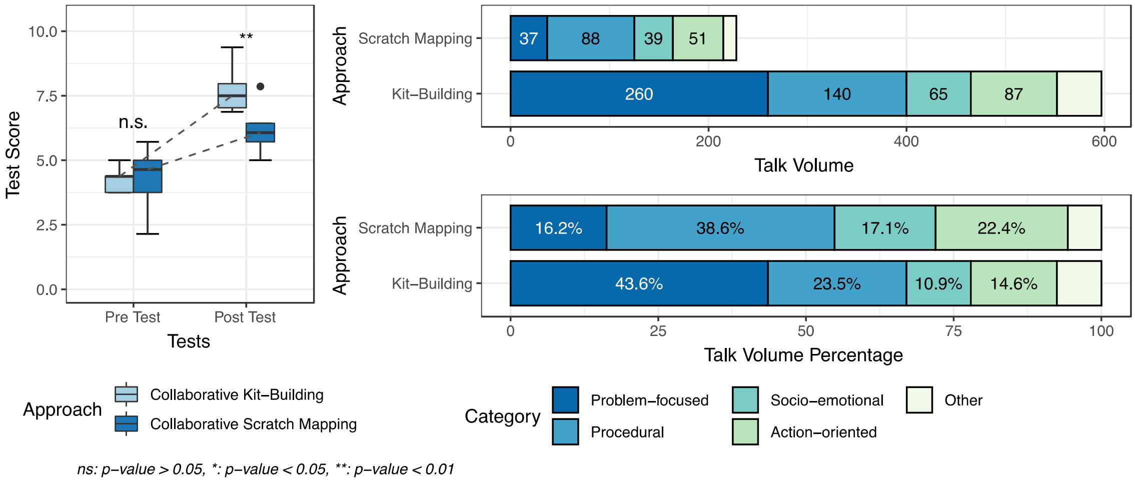


Figure 2. Comparison of Test Score and Talk Distribution.

4.2 *The Effect of Concept Map Activity and Discussion Towards Group Comprehension*

In addressing the second research question regarding students' activeness in the discussion that affects group comprehension, this study classifies the talks into act4teams categories and analyzes the talks after the collaboration. The group comprehension level after the collaboration was measured by post-test score. The Spearman correlation analysis at a 5% significance level between the volume of the talks to post-test score suggested no correlation for both kit-building and scratch mapping methods (*p-values* > 0.05). Furthermore, according to the Generalized Linear Model (GLM) analysis result as shown in Table 2, the students' comprehension was also not affected by the volume of discussion (*p-value* = 0.73017 > 0.05). Thus, an active discussion during collaboration does not necessarily reflect a higher comprehension of the learning topic. Students who actively discussed the concept maps with their partners have higher post-test scores than less active students. However, there were situations where students who comprehended the reading before collaboration talk and discussed less. Students who did not comprehend the topic may ask to get more information; thus, discuss more actively.

Table 2. *Generalized Linear Model Analysis Result for Post-Test*

	Estimate	Std. Error	t-value	*p-value*	Sig.
(Intercept) (Kit-Building)	7.79403	0.542588	14.365	6.14×10^{-11}	***
Method (Scratch Mapping)	-1.585756	0.459751	-3.449	0.00306	**
Total Talk Volume	-0.002831	0.008075	-0.351	0.73017	

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Analyzing the talks of act4teams categories with GLM also yielded similar results, as shown in Table 3. Neither of the volumes of the talks of each category influenced the student comprehension. The GLM analysis set the Kit-Building method as the reference variable of the linear model. The Estimate value for the Intercept represented the mean of the Kit-Building concept mapping method. According to the Estimate value of the model, the mean score of the Scratch Mapping method is lower than the reference Kit-Building method (-1.80488). In other words, the post-test score of the Kit-Building method has a higher mean score (7.85373) than the Scratch Mapping method (6.04885). Both concept mapping methods are shown to significantly influence students' group comprehension (*p-value* < 0.05).

Table 3. *Generalized Linear Model Analysis Result for Post-Test of Each Act4teams Talk Category*

	Estimate	Std. Error	t-value	p-value	Sig.
(Intercept) (Kit-Building)	7.85373	0.66802	11.757	2.67×10^{-8}	***
Method (Scratch Mapping)	-1.80488	0.61683	-2.926	0.0118	*
Problem-Focused Talk	-0.0217	0.02107	-1.03	0.322	
Procedural Talk	0.01115	0.03825	0.291	0.7753	
Socio-Emotional Talk	-0.03052	0.06982	-0.437	0.6692	
Action-Oriented Talk	0.03329	0.05289	0.629	0.54	
Other Talk	0.01957	0.07663	0.255	0.8024	

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

5. Limitation and Future Work

This study used a small number of participants in representing a collaboration group. Insignificant statistical results in this study may also be affected by the low number of samples be involved in the analysis. Therefore, it is difficult to interpret and generalize the result in a larger context due to a larger sampling error. Thus, evaluating the effects with larger samples is strongly suggested for future studies. This study also assumed that the participants could create a well-composed concept map after a short training session and use the collaboration system effectively.

Larger group size may affect how they collaborate and discuss while the concept mapping activity is carried out. Investigating how the students collaborate in a larger group with Kit-Build concept map is one interesting research topic to discuss and investigate in the near future. The act4teams coding scheme was used to categorize the talks during discussion. This study quantifies the discussion activeness based on the talk volume on each act4teams coding scheme rather than investigating the quality or how the students discuss with their collaboration partners.

6. Conclusion

This study suggested that concept mapping is a critical factor for successful collaborative learning in online settings. Using the Kit-Build concept map method helped students improve their comprehension better and encourage more active discussion than the traditional open-end concept mapping method in online collaborative learning with concept maps. According to the analysis result of this study, both concept mapping activities were suggested to influence student comprehension during collaboration. Students who actively discuss during collaboration could improve their comprehension even though the talks were not directly influencing their understanding. Nevertheless, Kit-Build concept map could help the students improve their comprehension better and further emphasized the benefit of Kit-Build as an alternative approach to support online collaborative learning.

Acknowledgements

This research is partially supported by JSPS KAKENHI Grant Numbers 19H04227.

References

Andoko, B. S., Hayashi, Y., Hirashima, T., & Asri, A. N. (2020). Improving English reading for EFL readers with reviewing kit-build concept map. *Research and Practice in Technology Enhanced Learning, 15*(1), 7. doi: 10.1186/s41039-020-00126-8

Chen, W., Allen, C., & Jonassen, D. (2018). Deeper learning in collaborative concept mapping: A mixed methods study of conflict resolution. *Computers in Human Behavior, 87*, 424– 435. doi: 10.1016/j.chb.2018.01.007

Dallimore, E., Hertenstein, J. H., & Platt, M. (2016). Creating a community of learning through classroom discussion: Student perceptions of the relationships among participation, learning, comfort and preparation. *Journal on Excellence in College Teaching, 27*(3), 137–171.

de Ries, K. E., Schaap, H., van Loon, A. M. M., Kral, M. M., & Meijer, P. C. (2021). A literature review of open-ended concept maps as a research instrument to study knowledge and learning. *Quality and Quantity*. doi: 10.1007/s11135-021-01113-x

Hirashima, T., Yamasaki, K., Fukuda, H., & Funaoi, H. (2011). Kit-build concept map for automatic diagnosis. In Proc. of Artificial Intelligence in Education (AIED) 2011, 466–468.

Hirashima, T., Yamasaki, K., Fukuda, H., & Funaoi, H. (2015). Framework of kit-build concept map for automatic diagnosis and its preliminary use. *Research and Practice in Technology Enhanced Learning, 10*(1), 17. doi: 10.1186/s41039-015-0018-9

Kauffeld, S., Lehmann-Willenbrock, N., & Meinecke, A. L. (2018). The advanced interaction analysis for teams (act4teams) coding scheme. In E. Brauner, M. Boos, & M. Kolbe (Eds.). In The cambridge handbook of group interaction analysis (pp. 422–431). Cambridge University Press. doi: 10.1017/9781316286302.022

Metcalf, S. J., Reilly, J. M., Kamarainen, A. M., King, J., Grotzer, T. A., & Dede, C. (2018). Supports for deeper learning of inquiry-based ecosystem science in virtual environments - Comparing virtual and physical concept mapping. *Computers in Human Behavior, 87*, 459–469. doi: 10.1016/j.chb.2018.03.018

Pailai, J., Wunnasri, W., Yoshida, K., Hayashi, Y., & Hirashima, T. (2017). The practical use of Kit-Build concept map on formative assessment. *Research and Practice in Technology Enhanced Learning, 12*(1). doi: 10.1186/s41039-017-0060-x

Pinandito, A., Hayashi, Y., & Hirashima, T. (2021). Online collaborative kit-build concept map: Learning effect and conversation analysis in collaborative learning of english as a foreign language reading comprehension. *IEICE Transactions on Information and Systems*, E104-D(7), 981-991.

Pinandito, A., Prasetya, D. D., Az-zahra, H., Wardhono, W., Hayashi, Y., & Hirashima, T. (2021). Design and development of online collaborative learning platform of kit-build concept map. *JITeCS (Journal of Information Technology and Computer Science), 6*(1), 50–65. doi: 10.25126/jitecs.202161294

Pinandito, A., Prasetya, D. D., Hayashi, Y., & Hirashima, T. (2021b). Semi-automatic concept map generation approach of web-based kit-build concept map authoring tool. *International Journal of Interactive Mobile Technologies (iJIM), 15*(08), 50–70.

Pinandito, A., Prasetya, D. D., Hayashi, Y., & Hirashima, T. (2021a). Design and development of semi-automatic concept map authoring support tool. *Research and Practice in Technology Enhanced Learning, 16*(1), 8. doi: 10.1186/s41039-021-00155-x

Prasetya, D. D., Hirashima, T., & Hayashi, Y. (2021). Comparing two extended concept mapping approaches to investigate the distribution of students' achievements. *IEICE Transactions on Information and Systems,* E104.D (2), 337-340. doi: 10.1587/transinf .2020EDL8073

Sugihara, K., Osada, T., Hirashima, T., Funaoi, H., & Nakata, S. (2012). Experimental evaluation of kit-build concept map for science classes in an elementary school. Proceedings of the 20th International Conference on Computers in Education, ICCE 2012, 17–24.

Tseng, S.-S. (2020). Using Concept Mapping Activities to Enhance Students' Critical Thinking Skills at a High School in Taiwan. *The Asia-Pacific Education Researcher, 29*(3), 249–256. doi: 10.1007/s40299-019-00474-0

van Blankenstein, F. M., Dolmans, D. H. J. M., van der Vleuten, C. P. M., & Schmidt, H. G. (2011). Which cognitive processes support learning during small-group discussion? The role of providing explanations and listening to others. *Instructional Science, 39*(2), 189–204. Retrieved from https://doi.org/10.1007/s11251-009-9124-7. doi: 10.1007/ s11251-009-9124-7

Wunnasri, W., Pailai, J., Hayashi, Y., & Hirashima, T. (2018). Reciprocal Kit-Build Concept Map: An Approach for Encouraging Pair Discussion to Share Each Other's Understanding. *IEICE Transactions on Information and Systems,* E101.D, 2356–2367. doi: 10.1587/transinf.2017EDP7420

Yoshida, K., Sugihara, K., Nino, Y., Shida, M., & Hirashima, T. (2013). Practical use of kit-build concept map system for formative assessment of learners' comprehension in a lecture. In Proceedings of the 21st international conference on computers in education. Indonesia: Asia-pacific society for computers in education (p. 906-915).

Theoretical and Practical Framework for a Multinational, Precollege, Peer Teaching Collaborative

Eric HAMILTON*, Danielle ESPINO & Seung LEE
Pepperdine University, USA
*eric.hamilton@pepperdine.edu

Abstract: This paper discusses a formal peer teaching collaborative network of middle and secondary school students designing and prototyping science, technology, engineering, and mathematics (STEM) projects that they share in synchronous virtual settings and asynchronous settings. Funded primarily by the US National Science Foundation, the effort has involved students from five continents. The aims of the collaborative, called the International Community for Collaborative Content Creation, or IC4, include fostering STEM skills and intercultural competence. The practical framework for IC4 is replicable. It aligns closely with theorizing around intercultural competence formation. The practical framework relies on interest-driven creator theory as a confirmatory guide for formulation of the projects that students undertake. The theoretical framework involves the construct of participatory teaching and involves quantitative ethnography, a methodology that relies on techniques from social network analysis and from discourse analytics to create visual and statistical models for phenomena traditionally expressed through case study. The paper includes discussion of how activity theory provides an important descriptive tool for explaining how collaboratives such as IC4 mediate the formation of academic and intercultural competencies.

Keywords: Quantitative ethnography, discourse analysis, computer-supported collaborative learning, ICT policy, global competencies, intercultural competence, virtual communication, activity theory, interest-driven creator theory

1. Introduction

This paper discusses a peer teaching collaborative that organizes teams of 12-19 year-old students who create and share science, technology, engineering, and mathematics (STEM) projects in synchronous virtual settings and asynchronous settings. Funded primarily by the US National Science Foundation, the effort has involved students from five continents. Aims of the collaborative, called the International Community for Collaborative Content Creation (IC4), include fostering STEM skills and intercultural competence. The practical framework for IC4 is replicable and tracks well with theorizing around recognized patterns in intercultural competence formation (Deardorff, 2006; Ramirez R, 2016). The practical framework relies on interest-driven creator theory (Chan et al., 2018) as a confirmatory guide for formulation of the projects that students undertake.

The theoretical framework involves the construct of participatory teaching as a means for adolescents to take on responsibility for learning about and then teaching STEM content to peers and to school teachers. The framework applies quantitative ethnography, a methodology that relies on techniques from social network analysis and from discourse analytics to create visual and statistical models for phenomena traditionally expressed through case study (Wooldridge, Carayon, Eagan, & Shaffer, 2018). The paper provides a sample visual model of how students develop collaborative competences. It then segues into a discussion of how activity theory (Greeno, 2016) provides an important descriptive tool for explaining how collaboratives such as IC4 mediate the formation of academic and intercultural competencies (Hamilton & Espino, 2020).

2. International Community for Collaborative Content Creation (IC4)

IC4 began in 2017 as a network of school-based and independent clubs from different countries. It has since included students from Kenya, Namibia, India, Brazil, Finland, the US, Iran, Singapore, the United Arab Emirates, Uzbekistan, Mexico, and Cameroon. The projects around which students collaborate have been conceptualized as *makerspaces*, which are often defined by physicality and by the opportunity they provide learners to manually experiment and construct artefacts that embody social cognition and obligate or spur intellectual growth (Peppler, Halverson, & Kafai, 2016). Among the most prominent makerspace domains are robotics, circuit board experiments, and 3D printing. A subset of the makerspace movement, though, involves digital activities. Among the most popular creative outlets are video making, games, coding, and commercial products such as Minecraft (Ripper and Secondo 2018).

This more expansive view of makerspaces encompasses the past decade's revolution in user-created digital media content. Because it takes place over internationally distributed virtual spaces, IC4 projects primarily (but not exclusively) fall into this subset of the makerspace movement. Students and teachers meet in Zoom videoconference sessions called global meetups, and in asynchronous Slack groups. The interests of participating students drive the selection and formulation of projects. This approach to marshalling the energy and enthusiasm of the participating students reflects the premise of interest-driven creator theory (Chan et al., 2018), which posits that the entrée for learner immersion requires leveraging the learner's motivations and interests, and do so through activities that furnish agency and a way to progress beyond surface-level interest to more sustainable and resilient engagement in that area.

Online global meetups have emerged as a key component in building the IC4 community. The opportunity for visual, synchronous communication both motivated and built social trust among the participants, increasing the depth of interactions with time and experience. As more meetups have taken place, a shared understanding of the culture and behavior at meetups has emerged (Hamilton and Owens 2018). This includes a shared understanding of the roles within the meetups, such as a facilitator that guides the conversation and presenter(s) who share their project. As students develop social trust and comfort in their makerspace culture, they are able to interact more openly across cultural and national boundaries.

Makerspaces provide a rich context not only for innovative student learning experience, but also for uncovering valuable insight for the effective design of future learning environments. Learning environments of the future will include routine and flexible, internet-mediated synchronous and asynchronous project collaboration (Dede 2010). Collaborations around making, or artifact creation in cross-cultural settings, obligate a variety of constructs and practices likely to alter and reshape future conceptions of learning. Among these constructs are three that IC4 emphasizes as an internationally distributed collaboration: social cognition, participatory teaching, and help-giving(Hamilton & Kallunki, 2020) These types of phenomena are likely to emerge in dynamic and highly positive forms in the future.

3. Assessing IC4 Participation through Quantitative Ethnography

From the outset of the IC4 network, it was clear that assessment of student participation would resist traditional approaches. The cross-cultural, age, prior knowledge, internet access, and school context differences each undermined evaluating experience through normative or standardized frameworks associated with academic achievement. The complexity of the challenge does not diminish the reality of academic achievement, but rather the inadequacy of available instruments to model or document achievement. Additionally, overarching interests by the research in fostering cross-cultural competence remain elusive to measure, in part for the same reason (differences across all baseline variables) and in part because the literature on building intercultural competence does not explore the context of adolescents collaborating across international boundaries or cultures through virtual tools (Hamilton & Espino, 2020). Though the pandemic is likely to address that gap, the field of adolescent international collaboration in academic contexts and its impact on intercultural competence has yet to take form.

It was in this context of seeking a means to model or explain IC4 experience and the growth it might stimulate that quantitative ethnography (QE) and the related analytic tool of epistemic network analysis (ENA) emerged as a promising methodology.

3.1 Quantitative Ethnography Operationalized by Discourse Analysis

A core premise of QE is that ethnographic study entails observation of socio-cultural patterns that shape our world, patterns that entail multiple layers of interconnections. Careful observation and articulation of socio-cultural patterns and the interconnections between them – ethnography - is certainly relevant for understanding and building policy and practice around innovation in digital media in learning and education, including changes we sought to foster in the IC4 ecosystem.

One of the most prominent objects of ethnographic observation in such as ecosystem at IC4 is discourse: how people communicate, in oral or written form, for example. Other tools include project artifacts, asynchronous versus synchronous balance, or visual or prosodic cues in conversations. Our focus on written and spoken discourse has allowed entrée to valuable analytic software tools only recently available in social science research. Such discourse tools enable analysis and visualization of large data sets by dint of increasing computational speed and storage. "Big data" discourse analytics provide a previously inaccessible yet powerful way to suggest, expose, or clarify ethnographic patterns whose articulation has traditionally been constrained to labor-intensive case studies. Analytics cannot replace ethnography – but can scaffold and give more finely grained resolution to ethnographic inquiry.

An essential step in applying discourse analytics to qualitative research more broadly is to define the mediating units of analysis. One such mediating unit in the domain of learning science research is called an *epistemic frame*. Epistemic Frame Theory (EFT) (Shaffer, 2006) treats a student's configuration of knowledge, skills, and experience, coupled with the individual's beliefs and self-efficacy, as a unit of analysis, or epistemic frame (Nash & Shaffer, 2012). Epistemic frames may be loosely compared to the construct of funds of knowledge (Moje et al., 2004)– i.e., the totality of unique experience, enculturation, beliefs, experiences, expectations, etc., that an individual brings to a social setting.) Epistemic network analysis (ENA) software developed under NSF funding (Marquart, 2018) provides analytic tools for graphing and interpreting discourse patterns. ENA software detects or enables visual interpretation of shifts in epistemic frames. It helps to measure whether or how IC4's objectives are reached. Sample graphs appear in Hamilton et al (2020).

4. IC4's Rationale: The Future Global Workforce

Considerations of the nature of the future STEM workforce contribute to shaping the rationale for IC4 and its applied research. Computer-supported collaborative learning research communities deeply understand that the global workforce will evolve rapidly and require preparation that differs significantly from current career planning patterns that characterize virtually every country, independent of national wealth or development profile. A census of career areas in 2030 and 2040 will bear little resemblance to such a census reflecting 2010 or 2020 occupations. The number of distinct career areas will expand. Many or most workforce position descriptions—or whatever "position description" as a construct evolves into—will have short lifespans. The workplace that middle and secondary school students of today will occupy in the future will continuously cycle in new ways of thinking and new tools. The pandemic-caused shutdown will not last, but the need to adjust to rapidly changing macro-conditions is likely to typify daily life in the future, in other ways and in different contexts. The rhetoric and literature that anticipates future technology-induced trends has coined terms such as Industry 4.0 or VUCA (volatility, uncertainty, complexity, and ambiguity) (Wallner et al., 2016). Another term, upskilling, has been familiar for many years to the labor market research community, and the literature on future workplace trends has incorporated it as well (Baldini, Botterman, Neisse, & Tallacchini, 2018). Upskilling represents a form of adaptive expertise (Baroody, 2003). Szalavetz (2019) and Müller et al. (2018) are among future workforce forecasters stressing the importance of technological upskilling or adaptive expertise to sustaining innovation. Expressions such as Industry 4.0, VUCA, upskilling, and others form a terminology or jargon for market observers and futurists. They point to the reality, that the future workforce will need to acquire complex competencies to take on relentless and unpredictable technological change.

What are these competencies? They include rapid conceptual migration (“rethinking”) and entrée to important new technologies as they come online, interpreting multiple design and use paradigms, and developing the agility to size up and rapidly master emerging technologies, when the velocity of technological innovation will be significantly higher than it is in the early 2020s. These must take form in humane ways that attend to fairness, helpfulness, and well-being of all. The IC4 project is intended to immerse its participants in experiences that will help those competencies and humane dispositions take form and flourish, and its participants build confidence for career decisions.

4.1 Intercultural Competencies among Adolescents in the Future Workforce

Among these global competencies for the future, one of the most prominent is intercultural competency. The field of intercultural competence includes multiple definitions, though they are not applied in the literature to the field of adolescents engaged in collaborative academic activity over virtual communication media. That field is only now in early formation. We have conjectured a cascading sequence or taxonomy of dispositions and skills that correspond to what may become a ubiquitous pattern in precollege education for building international collaboration programs.

The taxonomy involves matching aspects of intercultural competence with gradations of involvement. The taxonomy represents a work in progress for amalgamating research on workforce and tertiary settings with observations we have documented in IC4.

One overriding dynamic that appears conspicuously absent from workforce and tertiary setting research on intercultural competence involves intrinsic curiosity and joy in collaboration, dynamics prominent in the precollege setting of IC4 (Hamilton & Espino, 2020). Research on intercultural competence explicitly specializes in what workforce literature often refers to as soft skills and dispositions (Singh & Sharma, 2014). It is puzzling that research in this area does not more fully reference and build on one of the most prominent and energizing factors routinely evident in IC4: people are curious about other cultures, and if they can function in a non-threatening forum that establishes norms around respect, appreciation, and scientific wonder, they take active pleasure in working with collaborators from other cultures.

This has been a consistent finding in the IC4 network. This network also aligns with observations about intercultural competence involving collaborations between so-called global south and global north partners. Holmes (2017) notes that intercultural competence literature “… requires complementary research, education, and training that gives voice to those in the “global South” who may be marginalized, disenfranchised, poor, and exploited.” IC4 is in a unique position in that all of its core facilitators for global meetups represent low-income countries. Their voice in leadership neither highlights nor ignores north-south dynamics but rather deftly acknowledges those factors that may factor into any given collaboration.

4.2 Cultural Historical Activity Theory (CHAT)

ENA’s underlying principle that discourse reflects the enculturation and cross-enculturation processes of international virtual collaborations can be interpreted through cultural-historical activity theory (CHAT) (Greeno, 2016). A common premise of the learning sciences is that activity mediates learning (Radinsky & Gomez, 2000). Rather than preceding or preparing for activity, in other words, learning is embedded in activity systems. This is a key tenet of CHAT, and it corresponds to IC4’s emphasis on learning and problem solving while in collaborative makerspace-like activities. Various constructs of actors, rules and norms, instruments, community, and outcomes form the activity systems that mediate learning (Greeno, 2016). More importantly, treating internationally-distributed collaboration through a lens that focuses on cross-cultural, cross-national shared activity in a virtual space, in pursuit of outcomes (such as STEM challenges or other digital artifacts) changes terms by which school-age learners form perceptions of self and others in parts of the world that are remote to them. The virtual collaboration space, especially in synchronous video settings, enables visual communication with peers in other countries and cultures to take place from the familiarity of a student’s own culture and context (Hamilton & Kallunki, 2020).

Table 1. *Stages of Involvement in Collaborative Network and Corresponding Intercultural Competences Invoked and Developed*

Stage of Involvement	Relevant Disposition or Skill Exercised and Developed
Responds to recruitment opportunity	Intercultural and intellectual curiosity
Reviews informed consent with information about intercultural and international interactions	Priming for intercultural adaptation, flexibility, and negotiation
Meets peers and teachers from own and other countries in introductory phase	Intercultural curiosity (Hamilton & Espino, 2020)
Observes interactions and presentations by others, commenting superficially. Comments explicitly in group reflection	Empathy, sensitivity, respect, good will to listen and to understand across culture (Sun, 2014) Acknowledged pleasure and joy in cross-cultural interactions (Hamilton & Espino, 2020)
Formulates projects either individually or by responding to interests of others in the meetup	Developing flexibility and early intersubjectivity and shared meaning; cooperation (Stahl, 2016)
Carries out project individually or in collaboration with others	Intersubjectivity is further refined through aware of cultural nuance (Daly, 2016)
Formulates and shares presentation	Increased communication competence Competency to demonstrate across cultural mindsets (Chen, 2017)
Facilitates peers and teachers working on other projects	Interculturally aware, sensitive, and adroit (Chen & Starosta, 1996); Integrated and naturally flowing persona representing cognitive, affective, and behavioral (Genç, 2018; Zelenková, 2020)

The types of virtual collaboration activity system that take place within IC4 strongly appear to neutralize uncertainty, anxiety, or mistrust about those who live elsewhere by hybridizing physical presence - where the student is enculturated and at ease - with virtual presence in a collaborator's country and culture (Hamilton & Kallunki, 2020).

5. Forthcoming IC4 Directions: Larger Grain Projects and VR/Volumetric Interactions

IC4 will progress in two directions. One is that projects that students formulate and elect to undertake will take on a larger granularity. For example, one new initiative involves South America's Pantanal rainforest, and government-funded efforts by IC4 partners in Brazil, to position atmospheric condition sensors through Pantanal regions most vulnerable to destructive fires. In the initial year of implementation (2021), sensor data provide a corpus for neural network analysis and AI algorithm development, enabling predictions of fire likelihood. IC4 students in Brazil, the US, Mexico, and Sub-Sahara carry out parallel analyses with mirror data. This is one of several examples of IC4's evolution to projects of larger grain size, to help students build high-end media and AI competencies in international collaboration.

The second area of future development involves intensifying the hybridization of presence experience, and evaluating the conjecture that hybrid presence scaffolds the development of intercultural competence and learning, especially as it applies to trust building over systems that include video communication This will take place through developing shared virtual reality artifacts including volumetric presence (Cho, Kim, Lee, Ahn, & Han, 2020).

6. Conclusion

This phenomenon is familiar to adults accustomed to international virtual collaborations. In a world where strife and mistrust can germinate in part because of geographical or cultural differences, there is opportunity to invent fresh ways for school-age learners to understand those who do not live near them or do not live like them. This compelling dynamic applies both to geographic boundary-crossing and to cultural boundary-crossing that can occur within a country, a region, or even within a city (Hamilton & Kallunki, 2020). Displacing perceptions that originate in geographic, economic, cultural, or other differences with a productive and collaborative activity system as the primary basis for understanding those in other parts of the world is a different way to conceptualize intercultural competence.

Acknowledgements

The authors gratefully acknowledge the support of the US National Science Foundation and of Pepperdine University in sponsoring this research. This research does not reflect the views of the sponsoring organizations.

References

Baldini, G., Botterman, M., Neisse, R., & Tallacchini, M. (2018). Ethical design in the internet of things. *Science and engineering ethics, 24*(3), 905-925.

Baroody, A. J. (2003). The development of adaptive expertise and flexibility: The integration of conceptual and procedural knowledge. *The development of arithmetic concepts and skills: Constructing adaptive expertise*, 1-33.

Chan, T.-W., Looi, C.-K., Chen, W., Wong, L.-H., Chang, B., Liao, C. C. Y., . . . Ogata, H. (2018). Interest-driven creator theory: towards a theory of learning design for Asia in the twenty-first century. *Journal of Computers in Education, 5*(4), 435-461. doi:10.1007/s40692-018-0122-0

Chen, G.-M. (2017). 16 Issues in the conceptualization of intercultural communication competence. *Intercultural communication, 9*, 349.

Chen, G.-M., & Starosta, W. J. (1996). Intercultural communication competence: A synthesis. *Annals of the International Communication Association, 19*(1), 353-383.

Cho, S., Kim, S.-w., Lee, J., Ahn, J., & Han, J. (2020). *Effects of volumetric capture avatars on social presence in immersive virtual environments.* Paper presented at the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR).

Daly, A. (2016). Primary Intersubjectivity: Affective Reversibility, Empathy and the Primordial ‘We’. In *Merleau-Ponty and the Ethics of Intersubjectivity* (pp. 223-248). London: Palgrave Macmillan UK.

Deardorff, D. K. (2006). Identification and assessment of intercultural competence as a student outcome of internationalization. *Journal of studies in international education, 10*(3), 241-266.

Genç, G. (2018). Are Turkish EFL pre-service teachers ready to manage intercultural challenges? *Intercultural Education, 29*(2), 185-202.

Greeno, J. G. (2016). Cultural-Historical Activity Theory/Design-Based Research in Pasteur’s Quadrant. *Journal of the Learning Sciences, 25*(4), 634-639. doi:10.1080/10508406.2016.1221718

Hamilton, E., & Espino, D. (2020). Distributed Collaboration in STEM-Rich Project-Based Learning. https://repository.isls.org//handle/1/6849. In J. Roschelle, editor (Ed.), *Digital Promise Rapid Community Report Series, International Society of the Learning Sciences Repository*: Center for Innovative Research on Cyberlearning.

Hamilton, E., Espino, D., & Lee, S. (2020). *Frameworks And Affordances For Internationally Distributed Collaboration (IDC) Between School-Aged STME Learners.* Paper presented at the EpiSTEME8-Eighth International Conference to Review Research in Science, Technology and Mathematics Education, Tata Institute for Basic Research, Mumbai.

Hamilton, E., & Kallunki, V. (2020). Distributed Collaboration in Project-Based Learning and Its Assessment in Next-Generation Learning Environments. In E. A. Tatnall (Ed.), *Encyclopedia of Education and Information Technologies*: Springer.

Holmes, P. (2017). Intercultural communication in the global workplace, critical approaches. *The international encyclopedia of intercultural communication*, 1-16.

Marquart, C. L., Hinojosa, C., Swiecki, Z., Eagan, B., & Shaffer, D. W. . (2018). Epistemic Network Analysis (Version 1.7.0) [Software]. Available from http://app.epistemicnetwork.org.

Moje, E. B., Ciechanowski, K. M., Kramer, K., Ellis, L., Carrillo, R., & Collazo, T. (2004). Working toward third space in content area literacy: An examination of everyday funds of knowledge and discourse. *Reading research quarterly, 39*(1), 38-70.

Müller, J. M., Buliga, O., & Voigt, K.-I. (2018). Fortune favors the prepared: How SMEs approach business model innovations in Industry 4.0. *Technological Forecasting and Social Change, 132*, 2-17.

Nash, P., & Shaffer, D. W. (2012). *Epistemic youth development: Educational games as youth development activities*. Vancouver, BC, Canada.

Radinsky, J., & Gomez, L. M. (2000). *Making sense of complex data: a framework for studying students' development of reflective inquiry dispositions*: Citeseer.

Ramirez R, E. (2016). Impact on Intercultural Competence When Studying Abroad and the Moderating Role of Personality. *Journal of Teaching in International Business, 27*(2-3), 88-105. doi:10.1080/08975930.2016.1208784

Shaffer, D. (2006). Epistemic frames for epistemic games. *Computers and Education, 46*(3), 223-234.

Singh, M., & Sharma, M. K. (2014). Bridging the skills gap: Strategies and solutions. *IUP Journal of Soft Skills, 8*(1), 27.

Stahl, G. (2016). From intersubjectivity to group cognition. *Computer Supported Cooperative Work (CSCW), 25*(4-5), 355-384.

Szalavetz, A. (2019). Industry 4.0 and capability development in manufacturing subsidiaries. *Technological Forecasting and Social Change, 145*, 384-395.

Wallner, T., Wagner, G., Costa, Y., Pell, A., Lengauer, E., Halmerbauer, G., . . . Lienhardt, C. (2016). *Academic Education 4.0.* Paper presented at the International Conference on Education and New Developments.

Wooldridge, A. R., Carayon, P., Eagan, B. R., & Shaffer, D. W. (2018). Quantifying the qualitative with epistemic network analysis: A human factors case study of task-allocation communication in a primary care team. *IISE Transactions on Healthcare Systems Engineering, 8*(1), 72-82.

Zelenková, A. (2020). Using Cultural Taxonomies to Understand Intercultural Relations in Business. In *Exploring Business Language and Culture* (pp. 157-171): Springer.

Research on the Application of College Students' Online Learning Cognitive Engagement Evaluation

Yong-Hong WANG[a]* & Xiang-Chun HE[b]
[a]*College of Educational Technology, Northwest Normal University, China*
[b]*College of Educational Technology, Northwest Normal University, China*
*wangyh4437@nwnu.edu.cn

Abstract: College Students' online learning is gradually becoming more and more normalized. There is a correlation between learning engagement and learning quality. Cognitive engagement is important components of online learning engagement. Through literature research, expert consultation and analytic hierarchy process, this paper constructs the "online cognitive engagement evaluation index system of college students", which includes two first-class indicators, four second-class indicators, and determines the weight of each level of indicators. Through the design and development of the evaluation index system, based on the structural equation model of the observed variables on the corresponding latent variables of the factor load, the experimental class of college students online cognitive engagement was evaluated and analyzed, which provides reference for the development of online learning engagement evaluation of college students.

Keywords: Online learning, cognitive engagement, evaluation index system

1. Introduction

Learning engagement is an important factor to guarantee learning quality and influence learning performance, it is particularly important in the "online" situation. Learners' behavioral engagement is supported by their own internal psychological activities to achieve certain goals. Learners' cognitive engagement, such as strategy selection, monitoring and regulation, is an important factor affecting the quality of learning. This study starts with the connotation of online cognitive engagement, constructs the evaluation index system of College Students' online cognitive engagement, designs and develops evaluation tools, selects experimental objects for application, and provides reference for the research and practice of online learning cognitive engagement.

2. Cognitive Engagement in Online Learning

Cognitive engagement referred to the high degree of "participation" of learners' cognitive strategies and psychological resources (Heflin H, Shewmaker J & Nguyen J,2017), which was an element of online learning engagement. Most scholars believed that it mainly refers to learners' realization of their learning goals, selection of suitable methods and strategies in the process, and monitoring and regulation of their whole learning process (Lee,E. ,Pate,J.A.,& Cozart, D,2015). According to the composition of learning strategy coverage, Michael and others summarized learning strategies as cognitive strategies, metacognitive strategies and resource management strategies. Therefore, from the perspective of learning strategies, combined with the cognitive and metacognitive ideas proposed by Michael and others, the related research of scholars in cognitive engagement, and the cognitive characteristics of dialectical logical thinking and independent thinking of college students, this study defines "cognitive engagement in online learning" as: the learning strategies adopted by learners in order to achieve learning objectives in online learning environment Methods, skills and cognitive strategies to monitor and regulate the whole learning activities.

3. Determination and Evaluation of Evaluation Tools

The reliability coefficients of all dimensions were greater than 0.8, which indicated that the questionnaire had good reliability. The final "questionnaire of College Students' cognitive engagement in online learning" was shown in Table 1.

Table 1. *Questionnaire of College Students' Cognitive Engagement in Online Learning*

Evaluation dimension	Exogenous latent variable	Latent variable of internal cause	Topic content
cognitive engagement	cognitive strategy	rehearsal strategy	D2: After class, I will browse the platform resources to help complete the practical homework.
			D3: After class, I will review what I have learned on the platform in time to consolidate what I have learned.
			D4: I can relate the knowledge acquired on the platform to other disciplines.
		elaborative strategy	D5: When reviewing the contents I have learned online, I will often divide them into primary and secondary ones according to my own level and grasp the main points.
			D7: I will often summarize what I have learned into an outline to help me remember.
		organization strategy	D1: Before class, I will use the resources in the platform to preview what I want to learn.
			D6: I can summarize all kinds of resources in the platform into systematic knowledge.
	metacognitive strategy	planning strategy	D9: Usually, I will have a plan to review and practice the course content I have learned online.
			D10: Before the online test, I will make an effective review plan according to my actual situation.
			D12: I can carefully analyze the reasons for the mistakes in the online homework or test.
		monitoring strategy	D8: I will set my own online learning goals according to the guidance tasks.
			D13: I will review regularly according to the resources on the platform to help understand the relationship between knowledge points.
			D14: When I encounter difficulties in online learning, I will adjust my online learning plan and method in time.
		adjustment strategy	D11: When I study online, I often ask myself some questions to make sure I really understand what I have learned.
			D15: I will often evaluate and summarize the advantages and disadvantages of my online learning process.
			D16: I will often compare with other students to check their "online" learning methods and efficiency problems.

4. Structural Equation Model Analysis

Therefore, a second-order confirmatory factor analysis was conducted for the dimensions of "cognitive strategy” and “metacognitive strategy” with three first-order factors. Through calculation, the modified standardized parameter estimation model was shown in Figure 1.

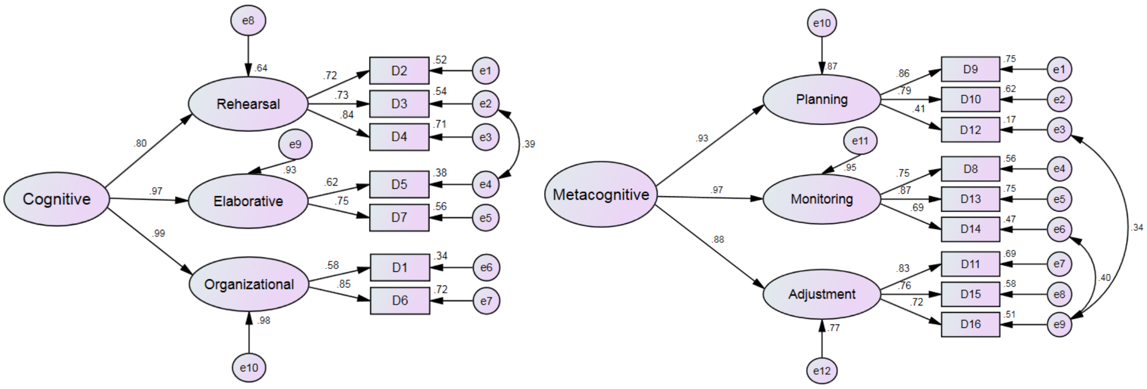


Figure 1. The Model Diagram of Standardized Parameter Estimation of Second-order Confirmatory Factor Analysis of “Cognitive Strategy” and “Metacognitive Strategy”.

5. Discussion

In the dimension of “cognitive strategy”, the cognitive engagement of college students was evaluated according to the factor load of each observation variable on its internal latent variable. In terms of "elaborative strategy", the vast majority of students in the two classes can often summarize what they have learned, which can effectively help to memorize knowledge points. In the aspect of "rehearsal strategy", the students of the two classes pay more attention to linking the learned knowledge with other subjects, which can better reflect the retelling strategy, so as to promote the online learners’ “cognitive input”. By browsing the platform resources to help complete the homework, and by reviewing the contents learned on the platform to consolidate the learned knowledge. However, there was a lack of summary and Reflection on the problems and deficiencies in the process of online learning.

6. Conclusion

Based on the established evaluation index system, the application research was carried out to explore the cognitive engagement of college students in online learning. The research found that: College students should be good at organizing all kinds of resources in the online platform into systematic knowledge, and pay attention to the organization strategy in the process of online learning; they should pay more attention to the connection and construction between knowledge and subject knowledge; they should summarize and reflect more on the problems and shortcomings in the process of online learning; only in this way can college students effectively promote the engagement of them in online cognition.

Acknowledgements

This research is supported by the 2021 Northwest Normal University Young Teachers' scientific research ability improvement plan project "Research on data literacy model construction and cultivation strategy of primary and secondary school teachers under the background of big data". Project No: NWNU-SKQN2021-29.

References

Heflin, H., Shewmaker, J., & Nguyen J. (2017). Impact of mobile technology on student attitudes, engagement, and learning. *Computers & Education*, 107, 91-99.

Lee, E., Pate, J. A., & Cozart, D. (2015). Autonomy support for online students. *Tech Trends*, (4), 54-61.

Integration of Programming-based Tasks into Mathematical Problem-based Learning

Zhihao CUI*, Oi-Lam NG & Morris S. Y. JONG
Department of Curriculum and Instruction, The Chinese University of Hong Kong, Hong Kong
*cuizhihao@link.cuhk.edu.hk

Abstract: In this paper, we presented four mathematical domain (arithmetic, random events and counting, number theory, and geometry) and corresponding tasks designed for several problem-based programming enrichment courses for middle school students. The courses aimed to integrate computing with mathematics to enhance mathematics teaching and learning. We examined the students' learning outcomes from each mathematical domain and task from three perspectives: cognitive, behavior, and affective with qualitative data. The results suggest that there existed affordances and challenges for learning both mathematics and programming in the tasks. We also identified two possible areas that contributed to the learning outcomes.

Keywords: Computational thinking, programming, mathematics education, problem solving

1. Introduction

Teaching with modern technologies is not only a currently popular topic but also a trend of future development. The scope of modern technologies used ranges from computers, tablets, 3D printing, and VR devices, in educational contexts ranging from college to secondary and elementary schools (Jong, 2015). Using programming and computational thinking to learning mathematical concepts can be traced back to Papert (1980) and further developed by Weintrop and his colleagues (2016), who illustrate the connection between mathematics and computational thinking. In our recent design-based research (Ng & Cui, 2020), we envisioned a computationally enhanced mathematics curriculum in local primary and secondary schools by exploring students' computational thinking development when engaging in mathematical problem solving with programming. We found that students' computational concepts, mathematical concepts, as well as problem-solving practices have been supported and developed through the designed tasks. However, as argued by Lockwood and De Chenne (2019), while programming seems to be effective in learning mathematics for certain topics, it cannot be concluded that it would be superior to paper and pencil. In addition, students were found to experience various challenges when solving mathematical problems in programming contexts (Cui & Ng, 2021).

To this end, there is still much room for investigation into the way computing can be integrated into mathematics education, especially in the K-12 context. One of the remaining questions in this regard was in exploring specific mathematical domains or topics that are suitable for integration with computing. In this paper, we explore four mathematical domains (arithmetic, random events and counting, number theory, and geometry) for different tasks used in a series of courses on mathematical problem-based learning via programming attended by a group of middle school students. In particular, we aim to address the research question, *what are the affordances for or barriers to learning both mathematics and computing with respect to the selected mathematical domains and tasks?*

2. Methodology

This study employed a research design of teaching experiments with qualitative data collection and analyses. The data reported in this paper was collected from several problem-based enrichment courses (named "digital making camps") designed for upper primary and lower secondary students as part of the study. We selected Scratch, a well-known and widely used block-based programming environment, as the computing tool from which mathematical problems are solved and presented by students.

Participants ranging from fifth- to eighth-grade (aged 10 to 14) were recruited from different primary and secondary schools in Hong Kong. We triangulated qualitative data in the form of students' artifact products, video and audio records, and self-reported surveys in reporting the results of the study.

3. Results

We present the results by the four mathematical domains (arithmetic, random events and counting, number theory, and geometry), respectively.

The observed learning outcomes for the tasks related to ***arithmetic*** were not as satisfactory as we had expected, despite that they were thought to be relatively simple compared to problems from other domains in a mathematical sense. The most common challenge was in the use of variables. Students had difficulties understanding the meaning of the variables they created or needed to create. Indeed, some students reported that the "bank balance problem" was the most difficult among all tasks. As commented by one student, "the numbers are too big, so a human brain is nearly impossible to do it, but I don't know how a computer thinks in this program, I use three days to complete that task […]". This suggested that the student saw the affordances of computing in dealing with large numbers; however, he struggled with solving the problem from a computational perspective.

In presenting problems with ***random events and counting*** (outcome space) in a programming environment, we observed that students had diverse performances in terms of cognitive and behavioral outcomes. This was related to the mathematics behind the problem, as we had intended that students would first program to experiment with random events using computer-generated randomized outcomes, and then observe the regularities of the frequency of outcomes to generate an outcome space. In the first part of the problem, the students achieved an average to a high degree in terms of both completion and quality. In comparison, the completion rate was significantly lower in the latter part of the problem. For example, we observe that many students wrongly used the "random" function in Scratch when dealing with the later part, which showed that the students remained at the level of experimenting with random events—a precursor to learning experimental probability. Nonetheless, for those students who did solve both parts of the problem, they demonstrated strong mathematical thinking and reasoning. One student simulated the dart-throwing situation 1000 times, and by observing the experimental outcomes of the dart's landing with the area of the dartboard, he inferred that the two quantities were proportional. Other students discovered that the outcomes ought to be symmetrical about the median when obtaining the six-dice sum (with equal likelihood). Regarding behavioral outcomes, we found that students were highly engaged in exploring and discussing the dice rolling problem. This was evident by their spontaneous discussion over the possible sums and most likely six-dice sum even before they began programming. In another incident, the students were taught to make the sprites move every time the random event generated a certain outcome. During the simulation of a large sample, the students betted on which sprite would move farther like a racing game with their pairs. It can be seen that the visual affordances of Scratch in supporting dynamic random events were positive to the students' learning experience in general.

In general, the quality and completion rate was unsatisfactory in the two tasks related to ***number theory***. Most students could program some artifacts, but few could solve the problems correctly with their programs. Regarding skill acquisition, some affordances of the programming environment were worth pointing out, including the possibility to test and debug one's solution to the problem. Since students were already familiar with the property of prime numbers, the students mainly engaged in testing their programs to make their "prime number detectors" work. The students knew that they needed to use conditions ("if… then") and repeat loops ("repeat… times" or "repeat until…"); however, they struggled to combine the two codes to repeatedly check the division statements, $N \div n$, where N is the given number and n is its possible factors. In such an open problem, the students had trouble ensuring the correctness of the program, especially without thoroughly and systematically testing. For example, one student claimed that his program worked, but in fact, it did not. It was because he had only used even numbers to test his problem. When the instructor asked him to test an odd composite number, the program detected the number as a prime. Therefore, it seemed that this task would be meaningful for developing testing and debugging practices in mathematical problem solving.

The students were most productive in programming ***geometry***-related solutions regardless of whether their solutions were correct or not. This can be evident in students programming various types of geometric shapes beyond what was required in the problem. For example, although we only instructed the students to draw triangles and squares, many students tried to extend their programs to drawing pentagon, hexagon, heptagon, etc. In doing so, they experienced functions and parameters, which were not typical for teaching and learning geometry with paper-and-pencil but were two computational concepts relevant for the lesson. That is, they named a function, "drawing polygon" with two parameters (i.e., number of sides), and then used the functions with different parameter inputs to draw various polygons. Creativity was observed when a student conveniently drew a 360-sided polygon using his program, which he called "a circle". Besides, using different and gradient changing colors was also common among students regardless of their level of programming skills. Where the target drawings were fractal geometry, i.e., Sierpinski Triangle, those students who failed to complete the tasks did create unique fractal geometry figures while exploring the solution. Without knowing why the program would output these particular figures, they continued to being engaged in generating them through trial and error. Many considered their programmable solutions "beautiful" figures, described as "flowers", "window grille", "carpet", "black holes", etc. The students were excited to see their programmed drawings, especially when their programs worked as designed. Most students reported in the post-camp surveys that the most enjoyable moment was to see their programs worked.

4. Discussion and conclusion

As informed by the findings, we suggest two areas in which programming and computational problem solving can afford enrichment to mathematics learning. The first is that the problem should be stated such that either the solution or the solution process is not immediately known. From this perspective, some mathematics-related problems were more effective when presented in a programming context. For example, the solution process (e.g., strategies for counting to 21), the solution itself (e.g., the prime detector for large number), or both (e.g., experimental probability and fractal geometry) were not known to the students immediately when it was first presented. This element of unknown provided opportunities for students to explore and inquire about new concepts, both in mathematics and programming. Secondly, computing with screen-based artifacts afforded dynamic visual representation and immediate feedback (such as movement of the sprite, outputting a certain number, a figure to be drawn, etc.), which has been shown to highly engage the students who participated in this study. As such, the students were more likely to continue regardless of the complexity and difficulty.

To conclude, this paper described and discussed some affordances and barriers to teaching and learning mathematics in computationally enhanced ways, drawing on selected mathematical domains and tasks. More research is warranted to further designing and studying learning materials for computationally enhanced mathematical teaching and learning.

References

Cui, Z., & Ng, O. (2021). The Interplay Between Mathematical and Computational Thinking in Primary School Students' Mathematical Problem-Solving Within a Programming Environment. *Journal of Educational Computing Research, 59*(5), 988-1012.

Jong, M. S. (2015). Does online game-based learning work in formal education at school? A case study of VISOLE. *Curriculum Journal, 26*(2), 249-267.

Lockwood, E., & De Chenne, A. (2019). Enriching students' combinatorial reasoning through the use of loops and conditional statements in Python. *International Journal of Research in Undergraduate Mathematics Education, 6*(3), 303-346.

Ng, O., & Cui, Z. (2020). Examining primary students' mathematical problem-solving in a programming context: towards computationally enhanced mathematics education. *ZDM – Mathematics Education, 53*(4), 847-860.

Papert, S. (1980). *Mindstorms, children, computers, and powerful ideas*. New York, NY: Basic Books.

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal Of Science Education and Technology, 25*(1), 127-147.

Web-Based Engineering Design Activity in Biology: An Assessment on the Demonstration of Higher-Order Thinking Skills

Ma. Andrea Claire CARVAJAL[a*], Catherine Genevieve LAGUNZAD[b] & Ma. Mercedes T. RODRIGO[c]
[abc]*Ateneo de Manila University, Philippines*
*andrea.carvajal@obf.ateneo.edu

Abstract: Senior High School (SHS) STEM curriculum in the Philippines is on its infancy stages which lacks technology-related and engineering-oriented courses. Along with the challenges in education sector brought by the COVID-19 pandemic and ineffective teaching of 21st century higher-order thinking skills in a Philippine setting, this study aims to utilize an engineering design activity to introduce engineering principles using technological tools and assess the presence or absence of higher-order thinking skills in the design solutions using an engineering design rubric. A design activity was developed by the researcher. The average and total scores of each group as evaluated by the subject teacher, students and researcher using the engineering design rubric to measure the demonstration of higher-order thinking skills in the outputs were analyzed. Based on the components assessed, problem-solving skills and critical thinking were demonstrated on the design solutions at competent and sophisticated level of performance. This shows that incorporating collaborative engineering design activities in an online setting allow the students to exhibit higher-order thinking skills.

Keywords: engineering design, assessment, higher-order thinking skills, senior high school

1. Introduction

Science, Technology Engineering and Mathematics (STEM) is one of the strands/tracks students can pursue in senior high school under the K-12 curriculum in the Philippines that was implemented in 2016 to become globally at par and ensure enhancement of skills. However, based on the subjects and competencies in the Philippine Senior High School (SHS) STEM curriculum set by the Department of Education (DepEd), it lacks the integration of technology-related or engineering-oriented courses and competencies (Arnilla 2018).

Problem-solving, innovation and design are the themes that are evident in the technology and engineering portion of STEM education (Hernandez et al. 2013). Engineering Design Process (EDP) is an efficient tool that teachers can use to introduce engineering principles to students, as well as develop higher-order thinking skills such as problem solving and critical thinking (Mangold et al. 2013; Ure 2012). Despite the many subjects in the basic education curriculum, higher order thinking skills such as critical thinking is not effectively taught in the Philippine setting (Marquez 2017). It is also supported by the Program for International Student Assessment (PISA) results reported in 2018--in which the examination uses the application of 21st century skills such as problem-solving, critical thinking and logical solutions, thus, incorporation of EDP in the curriculum could be a step on harnessing these higher-order thinking skills.

Along with changes on the educational settings brought by the COVID-19 pandemic which caused shifts from the traditional face-to-face class to online learning, utilizing EDP in an online setting where students have to make prototypes, could be a challenge. Nonetheless, studies show that through online support, students were able to familiarize the process of engineering design for analyzing open-ended design problems.

The aim of the study is to assess the demonstration of critical thinking and problem-solving skills through creating design solutions following the modified eight-steps of EDP which includes the iteration process of analyzing, selecting, creating, evaluating, and redesigning their solutions after identifying objectives and constraints as well as researching about their problem and communicating their final designs.

2. Methodology

Mixed method was utilized to analyze and gather data. For the of quantitative component, descriptive statistics was conducted through identifying the mean scores of the students on each criterion set in the engineering design rubrics, while observations from the students group interaction comprise the qualitative portion.

Engineering design activity in biology was developed by the researcher on the topic on energy transformation specifically ATP-ADP cycle, photosynthesis, and cellular respiration. 29 SHS STEM participants from a private school in Manila, Philippines were purposively selected, grouped and tasked to collaborate as a team and create an energy sustainable building design using the simulation software, Energy 3D following the modified eight-steps of EDP adapted from Mangold and Robinson (2013). Demonstration of higher-order thinking skills were evaluated by the subject teacher, students and researcher using the engineering design rubric under research and design, communication, and teamwork components using three-point rubric scale ranging from (3) sophisticated (2) competent and (1) not yet competent. The engineering design rubric was adapted from the project design rubric by the University of Pittsburgh. Each group's prototype design, written output as well as presentation were assessed. Due to the constraints of online set-up, teamwork component was evaluated through students' recorded discussions in creating their design outputs. Groups who did not conduct synchronous meetings submitted screenshots of their conversations in a messaging app.

3. Results

Based on the components assessed using the engineering design rubric through the mean scores, higher order thinking skills such as problem-solving skills and critical thinking were demonstrated on the design solutions at competent and sophisticated level of performance. Figure 1 shows a sample energy sustainable building design inspired by biological processes created by one group of students using the software Energy 3D.

The observations on the students' recordings and screenshots as they communicate with their teams show that similar routines were evident in all groups that contributed in the success of their design solutions. There was at least one member who led and initiates in the contribution of ideas, mostly males. It is then followed by the assignment of specific tasks, since not all of the members have computers/laptops in creating 3D designs, there is one member in the group who was assigned in creating the 3D design using the software Energy 3D. It was also observed in most groups that as a member propose a design idea it is followed by a confirmation from other members that they all agreed. Should the students find the proposed design solution lacking, members ask questions for clarifications and share improvements on the ideas. Members of some groups also search for design inspirations that were later modified based on their set goals and further evaluate the given suggestions. Some members also think about the possible limitations of the designs and apply their prior knowledge in creating solutions.

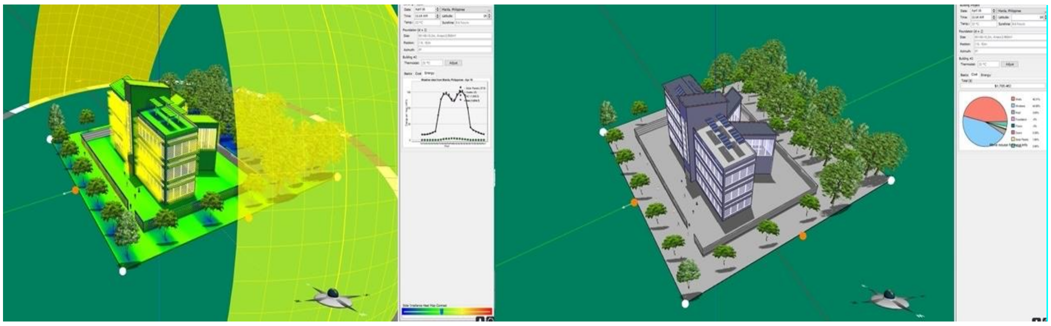


Figure 1. Sample Design by One Group using the Software Energy 3D.

4. Conclusion

Based on the components assessed using the engineering design rubric, it indicates that higher-order thinking skills such as problem-solving skills and critical thinking were demonstrated on the design solutions. Moreover, it shows that utilizing web-based engineering design activities in science concepts is one of the useful ways to introduce engineering and technology concepts to high school students.

It is recommended to impose true integrated STEM education in the Philippines by incorporating engineering and technology-based competencies on the science and mathematics courses. Through these, students will understand the relevance and connections of these fields as well as further improve necessary thinking skills.

Acknowledgements

I would like to give the highest glory, honor, and thanksgiving to our Dearest Jesus, who alone is the source of all strength and wisdom. I also want to express my deepest gratitude to my graduate institution, Ateneo de Manila University-School of Science and Engineering, which molded and equipped me with the knowledge and skills that I needed to accomplish this study. To my thesis advisers Dr. Cathy Lagunzad and Dr. Didith Rodrigo for their guidance and support. To the Department of Science and Technology-Science Education Institute, for the unwavering support to us scholars. And to our family and friends, who served as our inspiration in this endless pursuit of knowledge, thank you so much.

References

Arnilla, A. (2018). Possibilities and challenges for STEM methodology in the Public Senior high schools in the Philippines. Retrieved April 6 2021, from https://www.researchgate.net/publication/331980156_Possibilities_and_Challenges_for_STEM_Methodology_in_the_Public_Senior_High_Schools_in_the_Philippines.

Hernandez, P. R., Bodin, R., Elliott, J. W., Ibrahim, B., Rambo-Hernandez, K. E., Chen, T. W., & De Miranda, M. A. (2013). Connecting the STEM dots: Measuring the effect of an integrated engineering design intervention. *International Journal of Technology and Design Education, 24*(1), 107-120. doi:10.1007/s10798-013-9241-0.

Mangold J, Robinson S. (2013). The Engineering Design Process as a Problem Solving and Learning Tool in K-12 Classrooms. *ASEE PEER Document Repository. Sustainability and Manufacturing.* Retrieved June 13, 2020, from: https://peer.asee.org/22581.

Marquez, L. (2017). Philosophy in basic education: Towards the strengthening of the foundations of Philippine education. *Policy Futures in Education*, 147821031774365. doi:10.1177/1478210317743650.

Mcalpine I, Reidsema C. Allen B,.(2006). Educational Design and Online Support for an Innovative Project-based Course in Engineering Design. *Centre for Research on Computer Supported Learning and Cognition: Sydney University.* Retrieved June 13, 2020.

Ure, H. (2012). The effect of the engineering design process on the critical thinking skills of high school students. Retrieved June 13, 2020, from https://scholarsarchive.byu.edu/etd/3089.

Developing a Generic Skill Assessment System Using Rubric and Checklists

Makoto MIYAZAKI*, Hiroyoshi WATANABE, Mieko MASAKA & Kumiko TAKAI
Teikyo University, Japan
*miyazaki@lt-lab.teikyo-u.ac.jp

Abstract: In higher education, the development and assessment of generic skills are important issues, in addition to learning specialized knowledge. This paper clarifies the characteristics of generic skill assessment activities, and then describes the necessity of an information system for generic skill assessment in relation to Learning Management Systems (LMS) and e-portfolio systems, and finally, the requirements of such a system are proposed. The basic functions of the system that meets the proposed requirements were developed in accordance with the technical standards. More specifically, to manage assessment indexes such as rubrics, the database schema was designed with reference to IMS CASE. In addition, IMS LTI (Learning Tools Interoperability) was adopted, and the system was implemented as an LTI tool in order to be interoperable with educational systems such as LMS. The developed system was used in generic skill assessment activities at the Department of Information and Electronic Engineering, Teikyo University. The result of the usage demonstrated that the system was well-received by students, indicating that it can be fully used as an assessment system. A future issue will be to verify its usefulness as an assessment system by using the system over the long term.

Keywords: Generic skills, rubric, self-assessment system, LIT, CASE

1. Introduction

In recent years, educational reforms such as the introduction of active learning have been promoted in higher education. One of the objectives is to shift from the conventional "education that emphasizes knowledge" approach to "education that emphasizes both knowledge and generic skills." Generic skills are abilities and skills that can be used universally in any specialized field. Examples include communication skills, problem-solving skills, and the ability to work in a team. In order to provide education for the development of generic skills, it is necessary to assess them.

There are several methods to assess generic skills or 21st century skills (Geisinger, 2016). In the Programme for International Student Assessment (PISA) 2012, complex problem-solving skills were assessed by computerized simulations based on the multiple complex systems (MCS) approach (Herde, Wüstenberg, & Greiff, 2016). The Assessment and Teaching of 21st Century Skills (ATC21S) project took approach to assess collaborative problem-solving skills using log data on the real time collaboration task via an online assessment system (Care, Scoular, & Griffin, 2016). Although these are effective assessment methods, these focus on problem solving skills and difficult to deal with all generic skills in a unified manner.

Kawasaki, Kubota and Yajima (2020) adopted an objective test for checking generic skills called PROG for the assessment of generic skills. The PROG test consists of two parts: literacy and competencies. Of these, competencies determine the level of a test subject based on the similarity with the response results of working adults, but there is no certainty that the level determined in this way accurately represents the competency of the test subject. In other words, although PROG is an objective index, it is not always accurate. Generally, universities set their own educational goals based on their own diploma policy. These goals do not always match the assessment items of PROG. Therefore, PROG can be used as a reference when assessing generic skills, but it is not appropriate to assess by PROG alone.

Another approach is an intersubjective assessment using rubrics. For example, the VALUE Rubric Development Project (https://www.aacu.org/value/rubrics) by AAC&U in the United States developed rubrics to assess learning in undergraduate level education and are used in higher education.

Kyndt et al. (2014) created a questionnaire consisting of 44 items as a self-assessment instrument for generic working life competencies in vocational education. Sumaryati et al. (2019) developed a generic skills assessment index for use in self-assessments and peer assessments in accounting education. Khlaisang and Koraneekij (2019). developed an assessment system after defining assessment indexes for improving information literacy, media literacy, and ICT literacy.

The Department of Information and Electronic Engineering at Teikyo University (our department) has adopted PROG, an objective index, and CASEC, a diagnostic test for English communication skills. Our department decided to conduct a self-assessment using our own rubrics and checklists after referring to the results of these objective tests. This paper focuses on systems to perform a self-assessment using assessment indexes such as rubrics and checklists.

As the development of generic skills is important in higher education such as universities, the following is deemed necessary:

- The system should be easy to use in the process of assessing and developing generic skills in the university education curriculum.
- The system should be properly positioned in relation to LMS and e-portfolio systems as the educational information infrastructure of the university and should be interoperable with those educational systems.

However, there have been no existing examples of developing a generic skills assessment system from these points of view. Therefore, the objectives of this study are as follows:

To clarify the requirements that the generic skill assessment system used at the university must meet.

To implement the basic functions of the system according to the requirements and clarify its effectiveness.

In the following, Section 2 examines the general requirements of generic skill assessment activities based on the efforts of development and assessment of generic skills in our department and describes the necessity of a generic skills assessment system. Section 3 proposes a generic skill assessment system, and Section 4 describes the implementation of basic functions of the system. Section 5 describes the results of using the system in the assessment activities at our department, followed by discussions and a short summary.

2. Assessment of Generic skills

2.1 Example at Department of Information and Electronic Engineering, Teikyo University

In order to clarify the requirements for generic skill assessment activities, this section describes case studies of specific assessment efforts. Our department defined target generic skills to be assessed as shown in Table 1. Then, a generic skill rubric was created from the viewpoint of each item of the generic skills shown in Table 1. In addition, as a specific assessment index for each item, a checklist was created per item.

Table 1. *Target Generic Skills*

1. Information Literacy
2. Thinking and Problem-solving
3. Communication
   - 3.1 Written Communication
   - 3.2 Oral Communication
   - 3.3 Quantitative Literacy
   - 3.4 English Communication
4. Action and Learning
   - 4.1 Self-directed Learning
   - 4.2 Action
5. Team Work

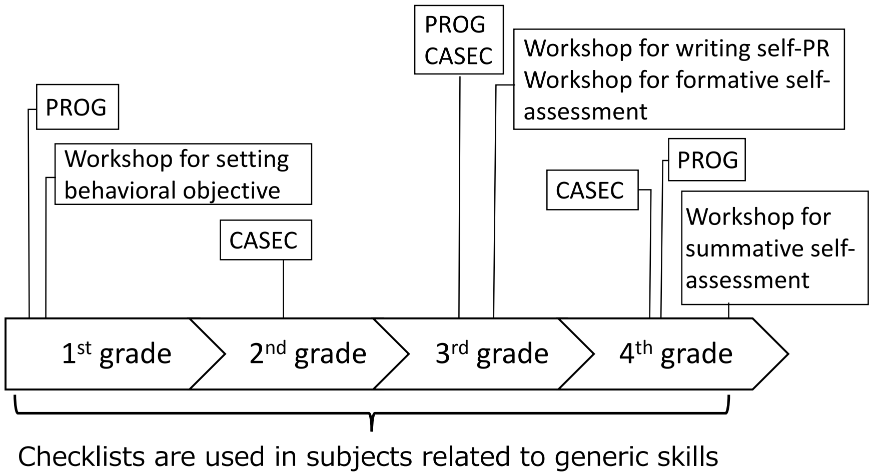

PROG
CASEC
Workshop for writing self-PR
Workshop for formative self-assessment
PROG
Workshop for setting behavioral objective
CASEC
CASEC
PROG
Workshop for summative self-assessment
1st grade
2nd grade
3rd grade
4th grade
Checklists are used in subjects related to generic skills

Figure 1. Assessment of Generic Skills in 4-Year-Educational Curriculum.

In many cases, specialized knowledge is taught in a specific subject, and it is thought to be appropriate to carry out an assessment in the subject. On the other hand, generic skills are not acquired in a specific subject alone, and so it is necessary to develop and assess them through the four-year educational curriculum at university. Thus, we decided to assess the generic skills as shown in Figure 1 during the four-year educational curriculum.

In Figure 1, the passage of time goes from left to right. The generic skill diagnostic test PROG and the English communication ability diagnostic test CASEC, which were adopted as objective indexes, were taken at the specified time. At the beginning of the second half of the third year and the end of the fourth year, an initiative called the "Achievement Confirmation Workshop" is conducted to self-assess generic skills overall. The final Achievement Confirmation Workshop for fourth-year students is somewhat of a summative assessment. Meanwhile, the self-assessment of each item is performed in the class related to each item of the generic skills. These assessments are a formative assessment, and also aim to make students aware of their generic skills.

2.2 Requirements for Generic skills Assessment Activities

From the efforts described in the previous section, it is considered that the generic skill assessment activities have the following characteristics:

(**Requirement 1**) It is necessary to carry out multiple assessment activities through the four-year educational curriculum at the university.

(**Requirement 2**) Basically, it is based on the student's self-assessment, but it should be assessed subjectively by using rubrics, etc., in addition to a subjective assessment.

Our department uses objective assessment indexes such as PROG together with the self-assessment, but in general, this is not an essential requirement.

(**Requirement 3)** In the assessment indexes such as rubrics and checklists, there are multiple assessment indexes that compose a parent-child relationship like overall generic skills (parent) and each item (child).

The generic skill rubrics we created and the checklists for each item of the rubrics have a parent-child relationship. In addition, Sumaryati et al. (2019) also developed a rubric made up of a parent-child relationship of (1) generic skills rubric and (2) measurement rubric of each attribute of generic skills, and it is considered to be a general requirement that the assessment index of generic skills has a hierarchy.

(**Requirement 4**) There are two types of assessment activities: one is to assess the generic skills of students as a whole, and the other is to assess specific items that make up the generic skills.

As generic skills are composed of various items, in subjects dealing with individual items, it is thought that there is a general need to assess only these items.

2.3 Necessity of Generic skills Assessment System

The utilization of an existing system is examined to carry out the assessment activities described in the previous section.

First, an LMS test function could be used for the self-assessment using rubrics and checklists. However, as the relevance of each test is not defined for LMS, it is not easy for faculty members and students to understand the assessment results by associating multiple items of generic skills or compare them with the past assessment results. Moreover, as LMS basically conducts management by subject (course), it is not suitable for operation beyond the scope of the subject. It is thought that the aforementioned Requirement 1 and Requirement 3 make it difficult to use LMS for the assessment of generic skills.

On the other hand, the use of an e-portfolio system is also conceivable. In the e-portfolio system, the main activity is to collect the output and compile it together as a portfolio, and then the portfolio is evaluated. Any e-portfolio system with rubric functions can be used to assess generic skills that meet the Requirements 1, 2, and 4 described above. However, we find no e-portfolio system that can manage an assessment index satisfying Requirement 3. Also, as generic skills consist of a relatively large number of assessment items, the burden on the students increases when trying to collect the output for all assessment items. In addition, the operation of assessment activities becomes complicated, and there is a greater burden on faculty members who run these assessment activities. Therefore, the following requirements are also important:

(**Requirement 5**) Assessment activities that do not require collecting output and evidence are needed.

Thus, as for showing evidence of assessment results and collecting related output, it is appropriate to selectively introduce them by considering trade-offs with operating costs.

Given the above, smooth operation will be difficult by utilizing the existing LMS and e-portfolio systems, so an information system specialized for the assessment of generic skills is required.

3. Proposal of Generic skill Assessment System

3.1 Functions of Generic skill Assessment System

Based on the requirements of the generic skill assessment activities described in the previous section, a system for the assessment of generic skills is proposed. It is desirable to limit the basic functions to those that are as simple as possible and yet are possible to perform assessment activities that meet the requirements, and extended functions for carrying out better assessment activities can be used selectively. The basic functions required for the assessment system are as follows:

- A function to manage assessment indexes such as rubrics and checklists. In management, the parent-child relationship of the assessment index can be handled like the overall generic skills versus each item.
- A function to manage the tasks of assessment activities according to the objectives, such as activities to assess the overall generic skills and activities to assess specific items.
- A function that allows students to make a self-assessment for a specified task and a function for students to visually browse their assessment results.
- A function that allows faculty members to visually browse the status of students' self-assessments. Also, a function to download the assessment results as a text file such as a csv file.
- A function that links with LMS and other educational information systems with regards to user ID and authentication.

The following functions can be considered as extensions to be used selectively:

- A function to manage and refer to the results of objective indexes such as PROG, which can be used as a reference in self-assessment.
- A function to present a portfolio that is the evidence of the assessment result linked thereto.
- A function to perform peer assessment and assessment by others such as teachers.
- A function to statistically analyze student assessment results.

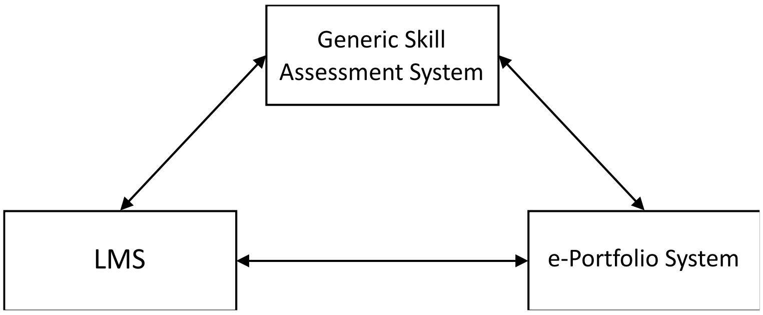


Figure 2. Relationship among LMS, e-Portfolio System and Generic Skill Assessment System.

3.2 Relevance to Existing Educational Systems

Figure 2 shows the relationship of the proposed generic skill assessment system to the LMS and e-portfolio system.

LMS is used on a day-to-day basis in class. As it is considered that assessment activities for specific items of generic skills are often carried out in class, it is desirable that the system has a linkage capability in which the assessment tasks of the generic skill assessment system can be seamlessly transferred from LMS. Given that LMS has been used in many universities in recent years, linkage between a generic skill assessment system and LMS should be implemented.

On the other hand, with the e-portfolio system, the aim is for students to develop by collecting and organizing the outputs for their learning goals and repeating activities to reflect on them, which leads to further learning. When collecting, organizing, and reflecting on output in association with class activities, it is desirable to be able to link with LMS. As it is conceivable to present a snapshot of the portfolio as evidence of the assessment results in the assessment of generic skills, the generic skill assessment system and the e-portfolio system should also be linked. However, compared to LMS, few universities use an e-portfolio system, so the generic skill assessment system and the e-portfolio system will be linked selectively.

Khlaisang and Koraneekij (2019) have made it possible to use an assessment system for improving information literacy, media literacy, and ICT literacy compatible with the Open edX platform or as a stand-alone system. Such linkage is important, and it is desirable to adopt a standardized method for linkage between systems. We, thus, decided to use IMS LTI (Learning Tools Interoperability, https://www.imsglobal.org/activity/learning- tools-interoperability).

4. Development of Basic System

4.1 Development Environment

In this paper, the basic functions of the proposed assessment system are implemented as an initial stage. The following was adopted as the development policy for the implementation:

- To actively comply with technical standards and standard specifications.
- To fully utilize open source software.

As the development framework, PHP's Laravel (https://laravel.com/) was employed. The system was an MVC model web application with a layered pattern of a presentation layer, business logic layer, and data access layer. Intuitive navigation was achieved by expressing information for operation and navigation with icons and pictograms. For this purpose, the front end of the view was designed using Bootstrap4 (https://getbootstrap.jp/), which is a web framework, and Font Awesome (https://fontawesome.com/), which is a web icon font.

4.2 Database Schema Design

For modeling of the rubric, the model and database schema were designed with reference to the rubric data model included in the standard specifications of IMS CASE (Competencies and Academic Standards Exchange, https://www.imsglobal.org/activity/case). The relationship among rubric data models in CASE is shown in Figure 3. It shows that the CFRubricCriterionLevel is a part of the CFRubricCriterion, and the CFRubricCriterion is a part of the CFRubric. Levels and points for each CFRubricCriterion were defined as follows: a rubric of academic standards and competencies is defined in CFRubric, rubric table rows in CFRubricCriterion, and rubric table columns in CFRubricCriterionLevel.

CASE data model names begin with the prefix CF. In consideration of readability in creating the model, different names were used for table names and column names as necessary like CFRubric being changed to Rubric.

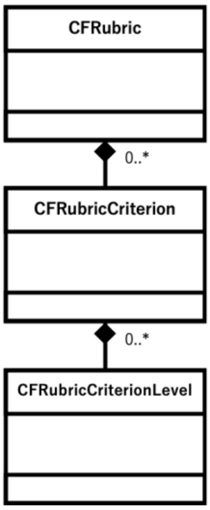
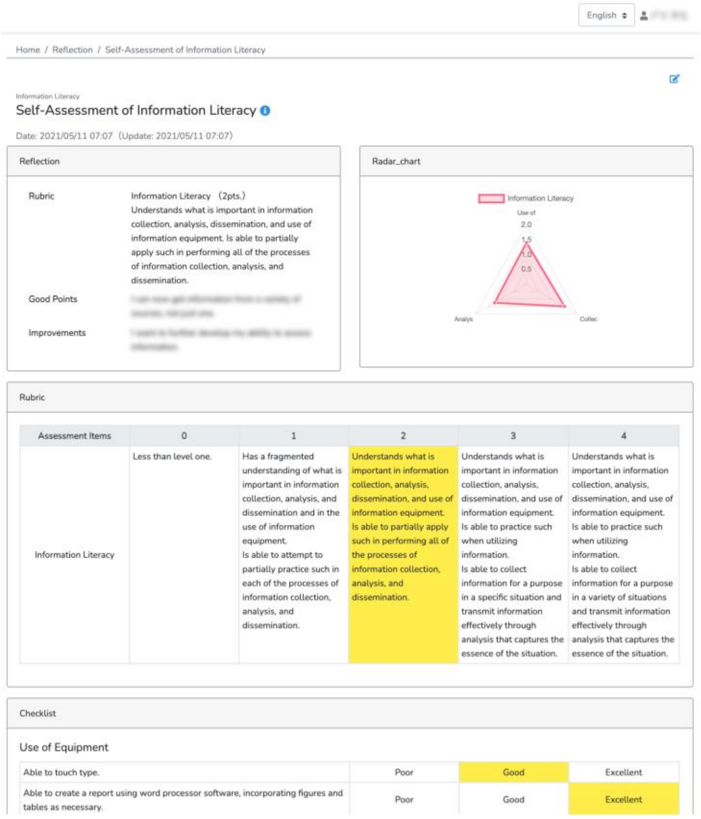


Figure 3. Relationship among Rubric Data Models in CASE.



[redacted]

Home / Reflection / Self-Assessment of Information Literacy

Information Literacy

Self-Assessment of Information Literacy

Date: 2021/05/11 07:07 (Update: 2021/05/11 07:07)

Reflection

Rubric	Information Literacy (2pts.) Understands what is important in information collection, analysis, dissemination, and use of information equipment. Is able to partially apply such in performing all of the processes of information collection, analysis, and dissemination.
Good Points	[redacted]
Improvements	[redacted]

Rubric

Assessment Items	0	1	2	3	4
Information Literacy	Less than level one.	Has a fragmented understanding of what is important in information collection, analysis, and dissemination and in the use of information equipment. Is able to attempt to partially practice such in each of the processes of information collection, analysis, and dissemination.	Understands what is important in information collection, analysis, dissemination, and use of information equipment. Is able to partially apply such in performing all of the processes of information collection, analysis, and dissemination.	Understands what is important in information collection, analysis, dissemination, and use of information equipment. Is able to practice such when utilizing information. Is able to collect information for a purpose in a specific situation and transmit information effectively through analysis that captures the essence of the situation.	Understands what is important in information collection, analysis, dissemination, and use of information equipment. Is able to practice such when utilizing information. Is able to collect information for a purpose in a variety of situations and transmit information effectively through analysis that captures the essence of the situation.

Checklist

Use of Equipment

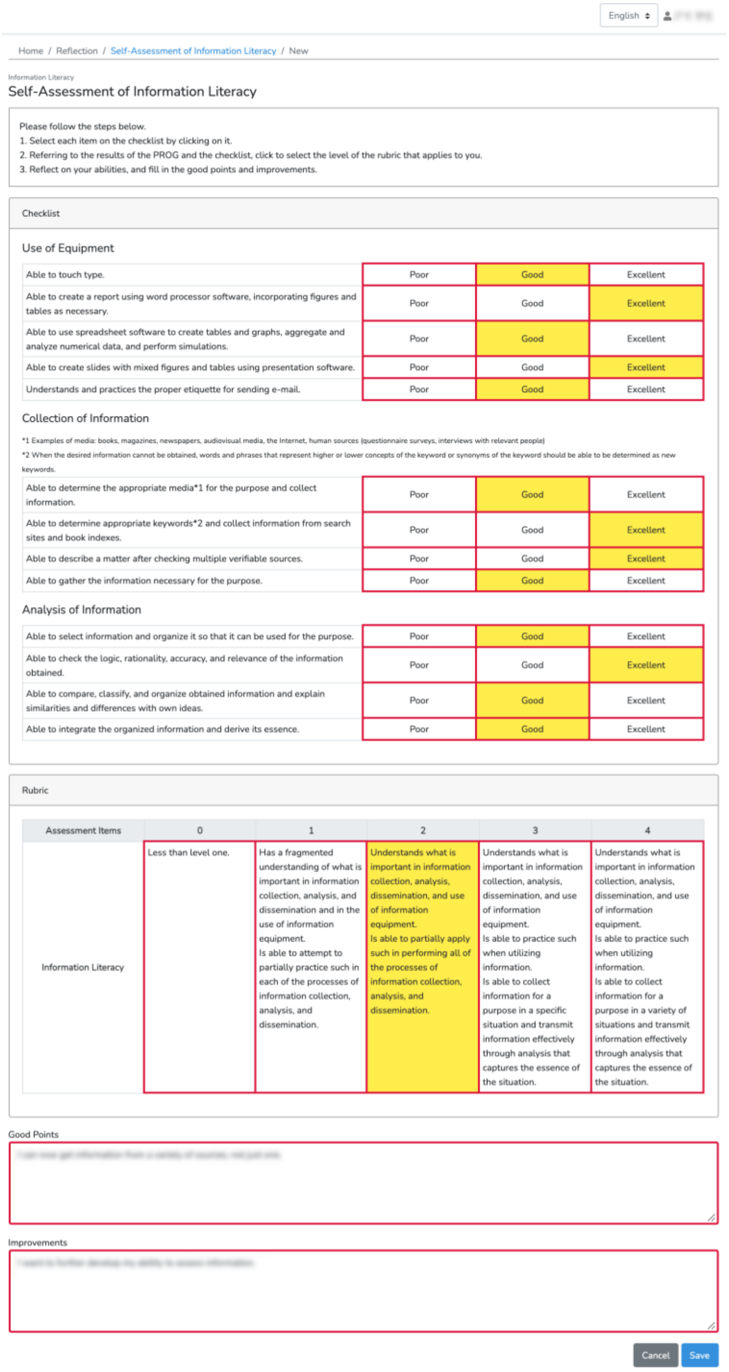
Able to touch type.	Poor	Good	Excellent
Able to create a report using word processor software, incorporating figures and tables as necessary.	Poor	Good	Excellent

Figure 4. An Example of a Self-assessment Result Screen for Students.

4.3 Implemented Functions

As a basic system, the basic functions described in section 3.1 were implemented. However, with regards to the "function to manage assessment indexes such as rubrics and checklists," the management itself was performed outside the system and imported into the system, but the system can manage the versions of rubrics and checklists and it is possible to handle the parent-child relationship of the assessment index. In addition, given that "assessment activities for the overall generic skills" can be replaced by "assessment activities for specific items," it was not implemented in the 1st version of the system.

Examples of the self-assessment result screen and student self-assessment input screen are shown in Figure 4 and Figure 5, respectively. Also, an example of the faculty's assessment activity summary screen is shown in Figure 6.



Home / Reflection / Self-Assessment of Information Literacy / New

Information Literacy

Self-Assessment of Information Literacy

Please follow the steps below.
1. Select each item on the checklist by clicking on it.
2. Referring to the results of the PROG and the checklist, click to select the level of the rubric that applies to you.
3. Reflect on your abilities, and fill in the good points and improvements.

Checklist

Use of Equipment

Able to touch type.	Poor	[✓ Good]	Excellent
Able to create a report using word processor software, incorporating figures and tables as necessary.	Poor	Good	[✓ Excellent]
Able to use spreadsheet software to create tables and graphs, aggregate and analyze numerical data, and perform simulations.	Poor	[✓ Good]	Excellent
Able to create slides with mixed figures and tables using presentation software.	Poor	Good	[✓ Excellent]
Understands and practices the proper etiquette for sending e-mail.	Poor	[✓ Good]	Excellent

Collection of Information

*1 Examples of media: books, magazines, newspapers, audiovisual media, the Internet, human sources (questionnaire surveys, interviews with relevant people)

*2 When the desired information cannot be obtained, words and phrases that represent higher or lower concepts of the keyword or synonyms of the keyword should be able to be determined as new keywords.

Able to determine the appropriate media*1 for the purpose and collect information.	Poor	[✓ Good]	Excellent
Able to determine appropriate keywords*2 and collect information from search sites and book indexes.	Poor	Good	[✓ Excellent]
Able to describe a matter after checking multiple verifiable sources.	Poor	Good	[✓ Excellent]
Able to gather the information necessary for the purpose.	Poor	[✓ Good]	Excellent

Analysis of Information

Able to select information and organize it so that it can be used for the purpose.	Poor	[✓ Good]	Excellent
Able to check the logic, rationality, accuracy, and relevance of the information obtained.	Poor	Good	[✓ Excellent]
Able to compare, classify, and organize obtained information and explain similarities and differences with own ideas.	Poor	[✓ Good]	Excellent
Able to integrate the organized information and derive its essence.	Poor	[✓ Good]	Excellent

Rubric

Assessment Items	0	1	2	3	4
Information Literacy	Less than level one.	Has a fragmented understanding of what is important in information collection, analysis, and dissemination and in the use of information equipment. Is able to attempt to partially practice such in each of the processes of information collection, analysis, and dissemination.	[✓ Understands what is important in information collection, analysis, dissemination, and use of information equipment. Is able to partially apply such in performing all of the processes of information collection, analysis, and dissemination.]	Understands what is important in information collection, analysis, dissemination, and use of information equipment. Is able to practice such when utilizing information. Is able to collect information for a purpose in a specific situation and transmit information effectively through analysis that captures the essence of the situation.	Understands what is important in information collection, analysis, dissemination, and use of information equipment. Is able to practice such when utilizing information. Is able to collect information for a purpose in a variety of situations and transmit information effectively through analysis that captures the essence of the situation.

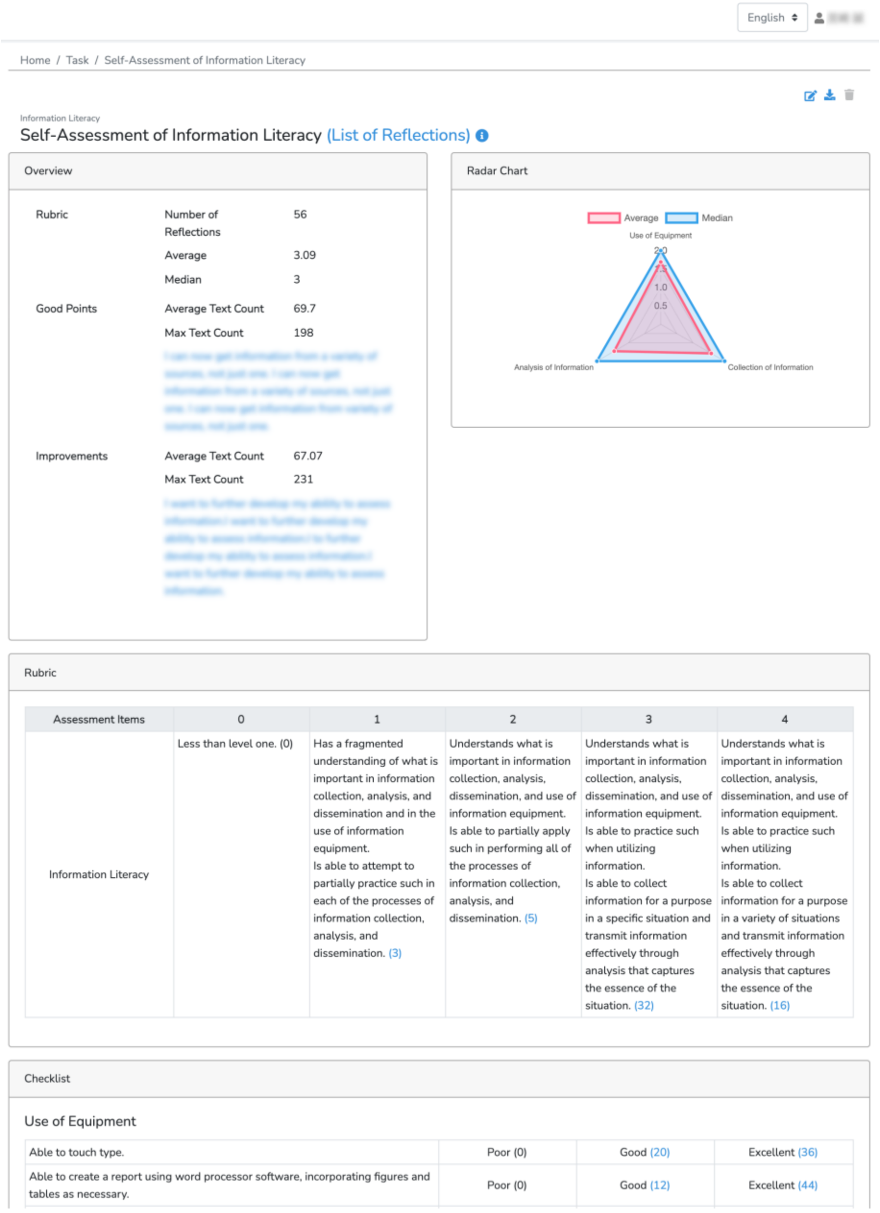
Good Points

[redacted]

Improvements

[redacted]

Figure 5. An Example of a Self-assessment Input Screen for Students.



Information Literacy

Self-Assessment of Information Literacy (List of Reflections)

Overview

Rubric	Number of Reflections	56
	Average	3.09
	Median	3
Good Points	Average Text Count	69.7
	Max Text Count	198
	[redacted]	
Improvements	Average Text Count	67.07
	Max Text Count	231
	[redacted]	

Radar Chart

Rubric

Assessment Items	0	1	2	3	4
Information Literacy	Less than level one. (0)	Has a fragmented understanding of what is important in information collection, analysis, and dissemination and in the use of information equipment. Is able to attempt to partially practice such in each of the processes of information collection, analysis, and dissemination. (3)	Understands what is important in information collection, analysis, dissemination, and use of information equipment. Is able to partially apply such in performing all of the processes of information collection, analysis, and dissemination. (5)	Understands what is important in information collection, analysis, dissemination, and use of information equipment. Is able to practice such when utilizing information. Is able to collect information for a purpose in a specific situation and transmit information effectively through analysis that captures the essence of the situation. (32)	Understands what is important in information collection, analysis, dissemination, and use of information equipment. Is able to practice such when utilizing information. Is able to collect information for a purpose in a variety of situations and transmit information effectively through analysis that captures the essence of the situation. (16)

Checklist

Use of Equipment

Able to touch type.	Poor (0)	Good (20)	Excellent (36)
Able to create a report using word processor software, incorporating figures and tables as necessary.	Poor (0)	Good (12)	Excellent (44)

Figure 6. An Example of an Assessment Activity Summary Screen for Faculty.

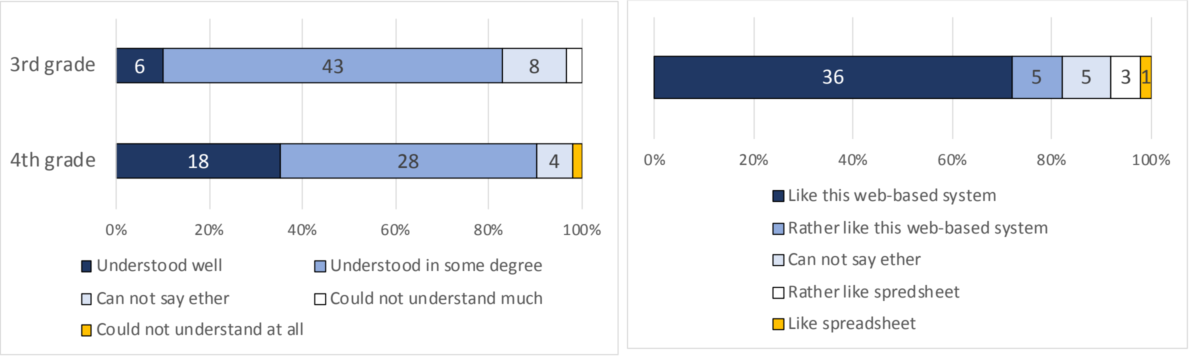
5. Evaluation of Basic System and Discussions

5.1 *Overview of Basic System Trial*

In order to verify the usefulness of the developed basic system, it was actually used on a trial basis at the Achievement Confirmation Workshop held by the fourth-year students in FY 2020 at the Department of Information and Electronic Engineering, Teikyo University.

At previous Achievement Confirmation Workshops, spreadsheet software was used for student self-assessment activities. More specifically, a worksheet for self-assessment was created as a spreadsheet, and the students downloaded the file, entered the self-assessment, and submitted it to the LMS. At the Achievement Confirmation Workshop for third-year students, the fourth-year students who experienced self-assessment using spreadsheets conducted a self-assessment using this system.

In the LMS Achievement Confirmation Workshop folder, a video explaining the significance and method of the self-assessment and a task of nine self-assessment items that make up the generic skills as an LTI link to this system were uploaded. The fourth-year students were notified that these activities be carried out from January 22 to February 5, 2021. After the completion of these activities, the students were asked to answer a questionnaire. Of the 59 fourth-year students, 51 completed all self-assessment activities and answered the questionnaire.



(a) Could you understand your generic skills?

(b) Which do you like this web-based system or spreadsheet?

Figure 7. Results of Questionnaire.

5.2 Questionnaire Results

Figure 7(a) shows the results of the questionnaire on the status of ascertaining the generic skills through assessment activities. Also shown are the results of a questionnaire similar to those conducted when they were third-year students (n = 59). The number of fourth-year students who answered "Understood well" increased compared to that when they were third-year students, and the total ratio of students who answered "Understood well" or "Understood in some degree" also increased. It is thought that the number of students who could understand well increased as the students matured year by year and deepened their understanding of generic skills. Although the use of this system cannot be said to be a factor in the increase, at the very least it demonstrates that the assessment activities using this system could be carried out without any problems.

Figure 7(b) shows the results of a questionnaire on the comparison between assessment activities using spreadsheets and those using this system. A total of 82% of students supported this system, with 72% of the students saying, "Like this web-based system" and a further 10% saying, "Rather like this web-based system." Many students also gave the following types of feedback supporting this system in the open comments field:

- The web-based system is easier to read and fill in.
- Except for the open comment part, answering was easy because I just had to click.
- I was able to input easily without feeling stress.
- The web-based system is much easier to use than the spreadsheets.
- I tried entering the information on an iPad and it was very easy to do.
- It was good that the information I entered could be confirmed as a diagram once I entered everything, and it was easy to understand my strong points and weak points.

On the other hand, the following comments were also found.

- It was easy to answer, but it was bothersome to have to return to the LMS each time I answered the checklist questions.
- It took extra effort because it was divided up for each item.

These are thought to indicate the need for a function that can handle the task of "assessment activities for the overall generic skills," which was not implemented in the 1st version of the system.

5.3 Discussions

Basic functions were implemented in the generic skill assessment system and the system was used by students in actual assessment activities. The results showed that it was able to be used without any problem for conducting assessment activities. From the results of the questionnaire, more than 80% of the students felt that it is better than the previous assessment activities using the conventional spreadsheet, and also from the open comments section, many students thought that this system is very convenient. On the other hand, one of the faculty members who is also one of the authors and was not involved in the implementation of the system but set up the assessment task on the LMS and checked the students' response status, stated that the system was easy to use from a faculty member's point of view as well. From these facts, it was made clear that this system is highly useful for both students and faculty members and can be fully used in generic skill assessment activities.

As the questionnaire results indicated the need for a function that can handle the task of "assessment activities for the overall generic skills," the function was implemented after the assessment by way of the trial run of the 1st version of the system.

This system was interoperable with LMS by way of LTI. The LMS of our university is Blackboard Learn, but we confirmed that it works properly with Moodle. It is thought that the system can be linked to any LMS that supports LTI. In addition, if the rubric is written in a CASE-compliant format, it can be imported into this system, and so the system can be used for assessment activities based on other assessment indexes. For example, generic skills can be assessed using rubrics different from those of our department. In addition, assessments other than generic skills are possible as long as they are based on rubrics.

The high interoperability of this system by adopting LTI and CASE enables the system to be used utilized in other institutions or different educational contexts. The skill development of students can depend on the curriculum and the instructional methods used in the courses. Also, methods or indexes of the assessment vary depending on the curriculum. Nevertheless, the system should be used as long as self-assessment activities using rubrics are conducted in the curriculum. We would like to investigate the effectiveness of the system in different educational contexts.

6. Conclusion

In this paper, after clarifying the requirements of a generic skill assessment system, a basic system that meets the said requirements was implemented and was verified of its usefulness by way of a trial run in the field. As a result, it was demonstrated that the system would work effectively in the assessment activities of generic skills at universities.

In future, while adding the extensions that can be used selectively as described in this paper, we would like to use this system over the long term and verify its usefulness in greater detail.

Acknowledgements

This work was supported by KAKENHI 17K12805 & 20K03075.

References

Care, E., Scoular, C. & Griffin, P. (2016). Assessment of Collaborative Problem Solving in Education Environments. *Applied Measurement in Education*, 29(4), 250-264.

Geisinger, K. F. (2016). 21st Century Skills: What Are They and How Do We Assess Them?. *Applied Measurement in Education*, 29(4), 245-249.

Herde, C. H., Wüstenberg, S. & Greiff, S. (2016). Assessment of Complex Problem Solving: What We Know and What We Don't Know. *Applied Measurement in Education*, 29(4), 265-277.

Kawasaki, K., Kubota Y. & Yajima, K. (2020). Proposal of a Generic Skills Improvement System in Which Students, Parents and Teachers Cooperate. *Proceedings of 5th International Conference on Information Technology*, InCIT2020, 7-11.

Khlaisang, J. & Koraneekij, P. (2019). Open Online Assessment Management System Platform and Instrument to Enhance the Information, Media, and ICT Literacy Skills of 21st Century Learners. *International Journal of Emerging Technologies in Learning (iJET)*, 14(7), 111-127.

Kyndt, E., Janssens, I., Coertjens, L., Gijbels, D., Donche V. & Petegem, P. V. (2014). Vocational Education Students' Generic Working Life Competencies: Developing a Self-Assessment Instrument. *Vocations and Learning*, 7, 365-392.

Sumaryati, S., Joyoatmojo, S., Wiryawan S. A. & Suryani, N. (2019). Alternative Assessment Instrument for Measuring Generic Skills of Accounting Education Students. *Proceedings of the 2nd International Conference on Education*, ICE 2019, 27-28.

Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Prior Knowledge on the Dynamics of Skill Acquisition Improves Deep Knowledge Tracing

Qiushi PAN [a*] **&Taro TEZUKA**[b**]

[a]*Graduate School of Comprehensive Human Sciences, University of Tsukuba, Japan*

[b]*Faculty of Library, Information and Media Science, University of Tsukuba, Japan*

* han.qiushipan@gmail.com

** tezuka@slis.tsukuba.ac.jp

Abstract: Knowledge tracing (KT) is the task of modeling how students' academic skills change over time. Given a sequence of a student's learning history, one goal of KT is to predict how well he/she will perform in the next interaction. Unlike in BKT (Bayesian knowledge tracing), the models in DKT (Deep knowledge tracing) cannot be improved simply by introducing elaborate prior knowledge about the task domain. Instead, we need to observe how trained models behave and identify their shortcomings. In this paper, we examine a problem in existing models that have not been discussed previously: the *inverted prediction problem*, in which the model occasionally gives predictions that are opposite to a student's actual performance development. Specifically, given an input sequence where a student has solved several problems correctly in a row, the model will occasionally estimate his/her skills to be lower than when he/she could not solve them. To tackle this problem, we propose *pre-training regularization*, which incorporates prior knowledge by supplying synthetic sequences to the neural network before training it with real data. We provide regular, simplistic synthetic data to a sequence-processing neural network as a specific implementation of pre-training regularization. This method solves the inverted prediction problem and improves the performance of the model in terms of AUC. We observed its effect qualitatively and introduced a quantitative measure to assess the improvement also. For ASSISTments 2009, ASSISTments 2015, and Statics 2011, improvements in AUC scores were 0.2 ~ 0.7 %, which are significant considering the scores are already high (around 70~80%). We developed an open-source framework for DKT with pre-training regularization. It also contains user-friendly hyperparameter optimization functionality.

Keywords: Knowledge tracing, educational data mining, recurrent neural networks, prior knowledge, regularization

1. Introduction

Online e-learning systems have rapidly grown in popularity in recent years. Students with different levels of understanding no longer need to be taught the same material in a classroom and can instead learn what is best suited for their knowledge level as predicted by a computer, based on their answers to past questions. By analyzing the accumulated big data, courses and assignments can be tailored for the best learning efficiency. Two examples of widely used e-learning platforms are intelligent tutoring systems (ITSs) and massive open online courses (MOOCs). With these systems, students are recommended which courses or lessons to take on the basis of their interaction with the system (Adamopoulos, 2013; Feng et al., 2009).

Knowledge tracing (KT) is the task of modeling students' academic abilities (Corbett and Anderson, 1994). Given a sequence of a student's learning history, the task is to predict the probability that he/she will answer the next question correctly. Bayesian knowledge tracing (BKT) handles this task by using a stochastic model. One of its merits is that the obtained parameters are easy to interpret. However, BKT requires human professionals to design the stochastic model carefully. Deep knowledge tracing (DKT) is an alternative approach that has become very popular in recent years (Piech et al., 2015). One of its strengths is that it does not require professionals to design its model parameters, as in BKT. It uses a recurrent neural network (RNN) (Zaremba et al., 2014) to capture the underlying structure of the students' understandings.

In this paper, we examine a problem of DKT that has been overlooked: the inverted prediction problem, where the student's performance is predicted as low even when he/she is consecutively answering questions correctly for some time. Based on this observation, we propose *pre-training regularization* and its unique implementation. Experiments showed that our proposed method mitigates the inverted prediction problem and leads to a higher AUC score. Furthermore, we developed a framework for training DKT based on our proposed method, together with easily accessible hyperparameter optimization functionality. Such functionality is especially valuable since it has not been available in previous implementations of DKT. The main contributions of this paper are as follows.

- We identify a unique problem in knowledge tracing, the inverted prediction problem, that we observe when making predictions using DKT. We propose a quantitative way of measuring its severity.

- We propose to use pre-training regularization and its specific implementation to overcome the inverted prediction problem. The proposed method also improved the AUC of the prediction results.

- We developed a framework for training DKT with pre-training regularization. Unlike existing DKT frameworks, our system contains user-friendly tools for hyperparameter optimization. The framework is open source and can be downloaded from the following repository.

https://github.com/qqhann/KnowledgeTracing

2. Related Work

2.1 Bayesian Knowledge Tracing

Many authors have pointed out that considering academic skills acquired by solving problems is crucial in knowledge tracing (Shepard, 1991; Resnick and Resnick, 1992; Cen et al., 2006). Bayesian knowledge tracing (BKT) was initially introduced by Corbett et al., 1994. It uses a binary stochastic variable for each knowledge concept (KC): understanding or not understanding. KC is a general term that represents a unit of knowledge, skill, or ability. In BKT, the probability of stochastic variables representing a KC is updated over time using hidden Markov models (HMMs). BKT represents the process of learning using probabilities of already knowing, guessing, acquiring, or slipping related to KCs. Based on these probabilities, BKT estimates a student's knowledge acquisition process over time. However, commonly used models in BKT are often oversimplified and rely on unrealistic assumptions, such as assuming that students will never forget what they have learned or that KCs are mutually independent and acquiring one KC will not affect the ease of acquiring other KCs.

Extensions of BKT include contextually modeling guessing and slipping (d. Baker et al., 2008), considering item difficulty (Pardos and Hefferman, 2011), and using student-specific parameters for individualization (Yudelson et al., 2013). There have also been proposals to reconsider the representation parameters of student abilities (Ritter et al, 2009). Wilson et al., 2016 showed that Hierarchical Item Response Theory (HIRT) and Temporal Item Response Theory (TIRT) could improve the performance if contextual information is used correctly.

2.2 Deep Knowledge Tracing

Deep knowledge tracing (DKT) was proposed by Piech et al., 2015. The model uses a recurrent neural network (RNN), especially a long short-term memory (LSTM) network. Unlike BKT that required domain specialists to adopt pedagogical knowledge into the model structure, DKT can learn it based on data. Although the trained parameters are difficult to interpret, it achieves a higher prediction accuracy. The RNN model forms the basis of DKT and is formulated as

$$\boldsymbol{h}_t = \tanh(\boldsymbol{W}_{hx}\boldsymbol{x}_t + \boldsymbol{W}_{hh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_h), \qquad \widehat{\boldsymbol{y}}_t = \boldsymbol{\sigma}(\boldsymbol{W}_{yh}\boldsymbol{h}_t + \boldsymbol{b}_y), \tag{1}$$

where the initial $\boldsymbol{h}_0$ is a randomly generated hidden state vector, $\boldsymbol{W}$ is the weight, and $\boldsymbol{b}$ is the bias. $\boldsymbol{\sigma}$ is a Q-dimensional-vector-valued function whose components are the sigmoid function. Q is the number of knowledge concepts (KCs). t is a time step. The student answers one question at each time step. $\boldsymbol{x}_t$ is encoded vector of (q_t, a_t) at time step t, where q_t is the KC that the question at time t belongs, and a_t is the correctness of the answer, which is 0 when the student answered the question incorrectly and 1 if he/she answered it correctly. Figure 1 shows the structure of the DKT model.

DKT has been actively investigated as a state-of-the-art approach to knowledge tracing (Xiong et al., 2016; Khajah et al., 2016; Yang et al., 2017; Sapountzi et al., 2018). In recent works, various shortcomings of DKT have been pointed out. Yeung and Yeung (2018) observed the *waviness* problem and *reconstruction* problem. Some extensions of DKT try to use more detailed study logs to achieve better results Zhang et al., 2017. Although more recent models have shown improvements over vanilla DKT, many still use an LSTM-based DKT as their base model. Therefore, we chose the vanilla DKT model as the baseline in this study and examined it in detail to identify fundamental issues.

Since DKT models are based on neural networks, it is not clear how one can incorporate prior knowledge about a student's learning process into the models. Unlike BKT, there is no systematic way of using a prior distribution. The aim of the current paper is to propose the use of pre-training as a way of introducing prior knowledge into DKT models. For each t, $\boldsymbol{x}_t = (q_t, a_t)$ is the 2-dimensional input vector, $\boldsymbol{h}_t$ is the hidden state vector, and $\hat{\boldsymbol{y}}_t$ is the Q-dimensional output vector, where Q is the number of knowledge concepts present in the dataset. The k-th component of $\hat{\boldsymbol{y}}_t$ represents the estimated probability that the student answers correctly at time $t+1$ to a question that belongs to knowledge concept k. The LSTM cell contains the weight $\boldsymbol{W}$ that is optimized by training. The LSTM cell and the weight is shared across time t.

3. Method

3.1 Inverted Prediction Problem

When examining the DKT model's prediction results, we often found that when a student is consecutively answering all questions correctly, his/her predicted performance (the probability that he/she will answer the next question belonging to the same knowledge concept correctly) is lower than that of students consecutively answering all questions incorrectly. We can roughly assume that once a knowledge concept (which represents a skill) is acquired, it is not lost easily. Therefore, the phenomenon of the decreasing predicted performance for questions belonging to a knowledge concept is counterintuitive. It suggests that the trained model does not represent the real mechanism of the student's learning process. We named this phenomenon, the inverted prediction problem.

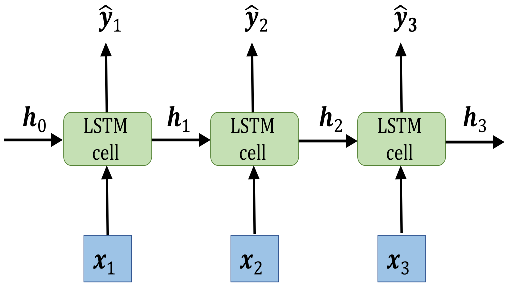


Figure 1. Schematic Diagram of DKT.

The opposite phenomenon maybe happening as well, where a DKT is unnecessarily optimistic and predicting that the student is more proficiently acquiring knowledge concepts than he/she really is. However, since such a phenomenon is harder to observe and quantify than the inverted prediction problem, we did not target it in the current paper.

Since the DKT model is a neural network, it is difficult to determine the cause of the phenomenon by analyzing its parameters. We chose to investigate it further by observing how the trained network responds to synthetic data. We created a synthetic sequence representing a student answering all questions correctly for a specific knowledge concept (KC). The length of the sequence is set to T. We call this sequence the *oracle student*, following Ding et al., 2019. The opposite is the *totally failing student* who answers all questions incorrectly.

To interpolate between the oracle and totally failing students, we created sequences where all questions were answered incorrectly up to a certain point, then the later questions are all answered correctly. These sequences represent students failing up to a certain point but then acquiring the necessary knowledge (skill) to solve problems and start answering correctly. Note that in real data, elements in a sequence belong to different KCs. In contrast, we used a synthetic sequence where all questions belong to a single KC q.

$$
\begin{gathered}
\boldsymbol{s}_0^q = ((q,0),(q,0),\dots,(q,0),(q,0),(q,0)) \\
\boldsymbol{s}_1^q = \big((q,0),(q,0),\dots,(q,0),(q,0),(q,1)\big) \\
\boldsymbol{s}_2^q = \big((q,0),(q,0),\dots,(q,0),(q,1),(q,1)\big) \\
\dots \\
\boldsymbol{s}_{T-1}^q = ((q,0),(q,1),\dots,(q,1),(q,1),(q,1)) \\
\boldsymbol{s}_T^q = ((q,1),(q,1),\dots,(q,1),(q,1),(q,1))
\end{gathered}
$$

We can formalize such sequences as:

$$
\begin{gathered}
\boldsymbol{s}_k^q = ((q,a_1),(q,a_2),\dots,(q,a_T)) \\
a_t = 1 \text{ if } t > T-k, \text{and } a_t = 0 \text{ otherwise,}
\end{gathered} \tag{2}
$$

where k is an integer in $[0,1,\dots,T]$, and q represents a KC. a_t is a Boolean value representing whether the student answered the question correctly or not at time t. In other words, when $a_t = 0$, the student answered a question belonging to q incorrectly at time t, and when $a_t = 1$, he/she answered it correctly. $\boldsymbol{s}_0^q$ represents a totally failing student, and $\boldsymbol{s}_T^q$ represents an oracle student. We trained the LSTM baseline model using real data from ASSISTments 2009 (Feng et al., 2009). We then added synthetic data to the model and examined the predicted performances of the students. Synthetic sequences consist of $\boldsymbol{s}_0^q$ to $\boldsymbol{s}_T^q$ for each KC. We used $T = 20$ in the following experiment.

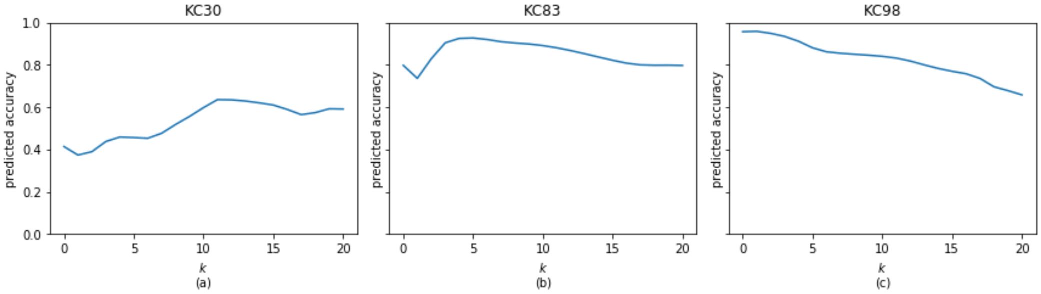


Figure 2. Predicted Performance Using Synthetic Sequences $\boldsymbol{s}_k^q$ Defined by Eq. 2.

The horizontal axes represent k and the vertical axes are the predicted performance $\hat{y}$. (a) is close to ideal dynamics since the predicted performance is nearly a monotonically increasing function of k. (c) is sub-optimal since the predicted performance is nearly a monotonically decreasing function of k. (b) is not as bad as (c) but is still sub-optimal.

In Figure 2, we use three values for KC q to explain the *inverted prediction problem*. The graph in (a) shows the result for KC with ID 30 ($q = 30$). When provided with the synthetic sequences $\boldsymbol{s}_0^q$ to $\boldsymbol{s}_T^q$ the model predicted that the longer the student correctly answered questions for the KC, the more likely he/she would be to answer the next question correctly. The result shown here matches our intuition about how students would perform. We call this property about students' predicted performance the *proportional prediction rule* since the predicted performance is proportional to how they performed earlier.

Unfortunately, the model did not learn this rule for all KCs. The graph shown in (c) is the worst-case example, where the predicted performance is a decreasing function of k. It means that when a student continued to answer questions correctly longer, the model predicted that his/her performance would be lower. Graph (b) is not as bad as (c) but is still sub-optimal because the predicted performance is a partially decreasing function of k. Although the model can distinguish between the oracle student and the failing student, it does not predict well for students performing in between. The inverted prediction problem shows one limitation of DKT when the available samples are limited. We use the problem to show how pretraining by prior knowledge can lead to better results than vanilla DKT.

3.2 Quantification of the Inverted Prediction Problem

To measure the severity of the inverted prediction problem, we propose two measures r_1 and r_2 formulated as

$$r_1 = \frac{1}{Q}\sum_{q=1}^{Q} \mathbf{1}_{\{\hat{y}(\boldsymbol{s}_T^q,q)>\hat{y}(\boldsymbol{s}_0^q,q)\}} \tag{3}$$

$$r_2 = \frac{1}{Q}\sum_{q=1}^{Q} NDCG\left(\hat{y}(\boldsymbol{s}_0^q,q),\ldots,\hat{y}(\boldsymbol{s}_T^q,q),(0,\ldots,T)\right). \tag{4}$$

r_1 gives a rough approximation of how well the model avoids the inverted prediction problem. Here $\mathbf{1}_{\boldsymbol{P}}$ is the characteristic function whose value is 1 when a proposition $\boldsymbol{P}$ is true, and 0 otherwise. q represents a KC, and Q is the total number of KCs. $\boldsymbol{s}_k^q$ is a synthetic sequence defined in Equation 2. $\hat{y}(\boldsymbol{x},q)$ is the output of the DKT model, representing the student's predicted performance when sequence $\boldsymbol{x}$ is the input for KC q. r_1 can be used to distinguish graphs (a) and (c) in Figure 2 but cannot be used distinguish (a) from (b).

Therefore, we propose r_2 that considers the overall dynamics of the predicted performance. Ideally, the predicted performance should be higher when the student answers more questions correctly. To quantify how well a model follows this pattern, we used the normalized discounted cumulative gain (NDCG) score, which is commonly used in evaluating ranking prediction (Wang et al., 2013). NDCG is a normalized measure of similarity between two orderings.

NDCG is calculated by dividing the discounted cumulative gain (DCG) by the ideal discounted cumulative gain (IDCG), that is, $NDCG = DCG/IDCG$.

Let n-dimensional vector $\boldsymbol{a}$ represent a sequence of numbers, $a_1,\ldots,a_n$, and n-dimensional vector $\boldsymbol{b}$ represent a sequence of numbers, $b_1,\ldots,b_n$. The numbers are not necessary in increasing or decreasing order. DCG is a similarity measure between two orderings defined as $\mathrm{DCG}(\boldsymbol{a},\boldsymbol{b}) = \sum_{i=1}^{n}\frac{\rho_i}{\log_2(i+1)}$ where ρ_i is the graded relevance score (Sawade et al., 2013) of $\boldsymbol{a}$ and $\boldsymbol{b}$ at position i.

$\mathrm{IDCG}(\boldsymbol{a},\boldsymbol{b})$ is equal to $\mathrm{DCG}(\boldsymbol{a},\boldsymbol{b})$ when $\boldsymbol{a}$ and $\boldsymbol{b}$ can be sorted using the same permutation ρ, that is, both $\rho(\boldsymbol{a})$ and $\rho(\boldsymbol{b})$ are sequences in increasing order; in other words, $\mathrm{IDCG}(\boldsymbol{a},\boldsymbol{b})$ equals to $\mathrm{DCG}(\boldsymbol{a},\boldsymbol{b})$ when $a_i \leq a_j$ if and only if $b_i \leq b_j$ for all pairs of indices i and j. In our case, such a case corresponds to $\hat{y}(\boldsymbol{s}_i^q,q) \leq \hat{y}\left(\boldsymbol{s}_j^q,q\right)$ for all $i \leq j$.

When the sequence $\hat{y}(\boldsymbol{s}_0^q,q),\ldots,\hat{y}(\boldsymbol{s}_T^q,q)$ is monotonically increasing for all q, $DCG\left(\hat{y}(\boldsymbol{s}_0^q,q),\ldots,\hat{y}(\boldsymbol{s}_T^q,q),(0,\ldots,T)\right)$ is equal to $IDCG\left(\hat{y}(\boldsymbol{s}_0^q,q),\ldots,\hat{y}(\boldsymbol{s}_T^q,q),(0,\ldots,T)\right)$, so r_2 takes the maximum value, which is 1. r_2 is large when DKT predicts that students who answered correctly longer will perform better for the next problem than students who answered correctly for a shorter period. r_2 is small when the opposite happens, that is, when the inverted prediction problem occurs.

3.3 Pre-training Regularization

To avoid the inverted prediction problem, we want to use prior knowledge that when a student continues to answer questions correctly for some time, his/her predicted performance will not suddenly drop. In transfer learning, the model is first pre-trained using a large dataset for a common task. The model is then further trained using a more specific dataset or a loss function representing a more specific task (Pan and Yang, 2010). We extend the idea of transfer learning by pre-training the model using synthetic data. In our proposed *pre-training regularization*, we generate synthetic data using simple rules that represent our prior knowledge about the process in which a student acquires knowledge.

We can gain some insights about pre-training regularization by comparing it with Bayesian modeling. In Bayesian modeling, the parameters of a prior distribution often correspond to pseudo-observations. For example, consider using a multinomial distribution as a generative model $p(\boldsymbol{x}|\boldsymbol{\mu})$

where $\boldsymbol{x}$ and $\boldsymbol{\mu}$ are d-dimensional vectors. Its conjugate prior $p(\boldsymbol{\mu}|\boldsymbol{\alpha})$ is the Dirichlet distribution. The posterior distribution is

$$p(\boldsymbol{\mu}|\boldsymbol{x},\boldsymbol{\alpha}) = \frac{p(\boldsymbol{x}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha})}{p(\boldsymbol{x}|\boldsymbol{\alpha})} = \left(\prod_{h=1}^{d} \mu_h^{x_h}\right)\left(\frac{1}{B(\boldsymbol{\alpha})}\prod_{h=1}^{d} \mu_h^{\alpha_h - 1}\right)\frac{1}{p(\boldsymbol{x}|\boldsymbol{\alpha})} = \frac{1}{p(\boldsymbol{x}|\boldsymbol{\alpha})B(\boldsymbol{\alpha})}\prod_{h=1}^{d} \mu_h^{x_h+\alpha_h-1}$$

where $B(\boldsymbol{\alpha})$ is the Beta function. The maximum a posteriori (MAP) estimate of $\boldsymbol{\mu}$ maximizes $p(\boldsymbol{\mu}|\boldsymbol{x},\boldsymbol{\alpha})$. Using the method of Lagrange multipliers to fulfill the constraint $\sum_{h=1}^{d} \mu_h = 1$, the estimate for μ_h can be obtained by

$$\hat{\mu}_h = \frac{x_h + \alpha_h - 1}{\sum_{u=1}^{d}(x_u + \alpha_u - 1)}.$$

In the above equation, the hyper-parameter α_h is added to the observation x_h which represents the number of occurrences of the event indexed by h. Incrementing α_h by one corresponds to increasing x_h by one, which means observing one extra occurrence of the event h. For this reason, α_h is called a *pseudo-observation*. Unlike real observations, pseudo-observations can take any non-negative real value. This correspondence between a hyper-parameter and a pseudo-observation is commonly seen in many Bayesian models. Since pre-training by synthetic data corresponds to adding extra observations to real data, it amounts to adding pseudo-observations to real data. In this way, pre-training corresponds to using prior knowledge. As DKT is based on deep learning rather than Bayesian modeling, its hidden layers and nodes can be considered as latent stochastic variables, and the probabilistic interpretation of deep learning can thus be a powerful analysis tool.

We propose to use a specific example of pre-training regularization as a remedy to the inverted prediction problem. We pre-trained DKT using monotonic synthetic sequences. Specifically, we generated synthetic sequences $\boldsymbol{s}_0^q$, $\boldsymbol{s}_1^q$, …, $\boldsymbol{s}_T^q$ (defined by Equation 2) for all KCs and used them to train the DKT network before training it with real data.

Another way to look at pre-training regularization is that it brings the output closer to a monotonically non-decreasing step function expressed as $\boldsymbol{s}_0^q$ to $\boldsymbol{s}_T^q$. For each KC, the model is pre-trained using sequences $\boldsymbol{s}_0^q$ to $\boldsymbol{s}_T^q$ repeatedly. We trained the model using synthetic data for all KCs for a certain number of epochs (e.g., 10 or 150) before training it with real data. By providing the network with sequences having a simple structure in pre-training, we expect the model to prefer simple dynamics when making predictions. Since this method does not require designing a prior distribution, it can be adapted to any model with ease. No domain knowledge is required, except the assumption that once a student acquires a knowledge concept, it is unlikely that he/she will lose it within a short time span.

4. Experiments

4.1 Datasets

ASSISTments 2009-2010: In the experiment, we used the ASSISTments *skill builder* dataset 2009-2010 (Feng et al., 2009). ASSISTments is an online tutor system that helps teachers access students' assessment information while studying math on the system. The dataset contains records of 4,417 students solving 328,291 problems. Each problem is manually tagged with a single skill concept required to solve it, out of 110 skill concepts in total.

ASSISTments 2015: ASSISTments 2015 is a dataset of 19,840 students solving 683,801 problems from 100 KCs.

Simulated-5: Simulated-5 is a dataset consisting of answer logs from simulated students based on the Item Response Theory (IRT) (Wilson et al., 2016). It contains records from 2,000 simulated students solving problems from 50 KCs.

Statics 2011: Statics 2011 is a dataset from students taking an engineering statics course. It has data of 333 students' 189,297 interactions on 1,223 KCs.

4.2 Results

Experimental results are summarized in Table 1. For comparison, we reconstructed the DKT model proposed by Piech et al., 2015. The baseline DKT is set up based on the previous research. We used a hidden dimension size of 200 and a batch size of 128. We set the learning rate to 0.1. We optimized hyperparameters using 5-fold cross-validation for ASSISTments 2009, ASSISTments 2015, and Statics 2011. We did not use cross-validation for Simulated-5 since we wanted to test it using the same condition as Piech et al., 2015. Cross-validation was carried out at the student level, that is, there is no student appearing in multiple folds.We searched for the best number of iterations for pre-training using validation data. We found that in most cases, 10 iterations were sufficient. We also tested 150 iterations to show that if there is a significant improvement by further pre-training.

Table 1 shows that the values of r_2 consistently increased with pre-training, suggesting it to be a better indicator than r_1. Pre-training improved AUC for most datasets. Pre 0 is without pre-training, pre 10 is with pre-training for 10 epochs, and pre 150 is with pre-training for 150 epochs. The values of r_1 are followed by the number of KCs in the corresponding dataset, which is the maximum possible value that r_1 can take. r_2 shows the mean value with standard deviation. Numbers in bold font are results that outperformed the baseline (DKT pre 0). The distribution of values for AUC comes from cross-validation. The distribution of values for r_2 represents the standard deviation of the terms in the sum in Equation 2.

Table 1. *Comparison of Area-under-the-Curve (AUC), r_1 and r_2*

Dataset	Model	Test AUC	r_1	r_2
ASSISTments 2009	DKT pre 0	0.802399 ± 0.002135	91 / 110	0.8858 ± 0.1210
ASSISTments 2009	DKT pre 10	**0.802763 ± 0.001107**	105 / 110	**0.9162 ± 0.1031**
ASSISTments 2009	DKT pre 150	**0.805398 ± 0.000801**	**106 / 110**	**0.8974 ± 0.1125**
ASSISTments 2015	DKT pre 0	0.703117 ± 0.001366	**99 / 100**	0.9243 ± 0.0914
ASSISTments 2015	DKT pre 10	**0.703355 ± 0.000890**	**99 / 100**	0.9211 ± 0.0967
ASSISTments 2015	DKT pre 150	**0.705414 ± 0.001263**	95 / 100	**0.9278 ± 0.0926**
Simulated-5	DKT pre 0	**0.777116**	26 / 50	0.7923 ± 0.1306
Simulated-5	DKT pre 10	0.776778	**44 / 50**	**0.8948 ± 0.1156**
Simulated-5	DKT pre 150	0.773313	**46 / 50**	**0.8908 ± 0.1139**
Statics 2011	DKT pre 0	0.787167 ± 0.000661	639 / 1223	0.7788 ± 0.1496
Statics 2011	DKT pre 10	**0.794149 ± 0.003777**	**934 / 1223**	**0.8855 ± 0.1267**
Statics 2011	DKT pre 150	**0.792491 ± 0.000426**	**716 / 1223**	**0.8095 ± 0.1582**

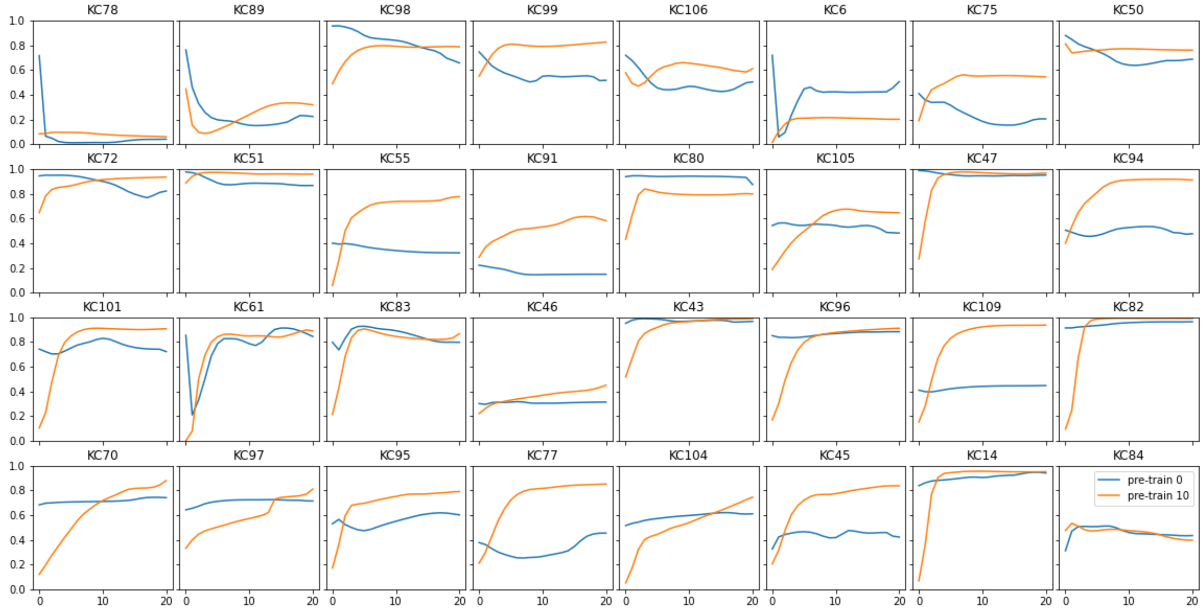


Figure 3. Prediction accuracy for KCs in ASSISTments 2009 when pre-trained using synthetic $\boldsymbol{s}_t^q$ defined in Equation 2. With pre-training, the inverted prediction problem occurred less. Blue is the baseline and orange is DKT with pre-training for 10 epochs. KCs are sorted by $\hat{y}(\boldsymbol{s}_T^q) - \hat{y}(\boldsymbol{s}_0^q)$. Vertical axis is the prediction accuracy for the same KC as used for the synthetic input. Horizontal axis represents knowledge concept k.

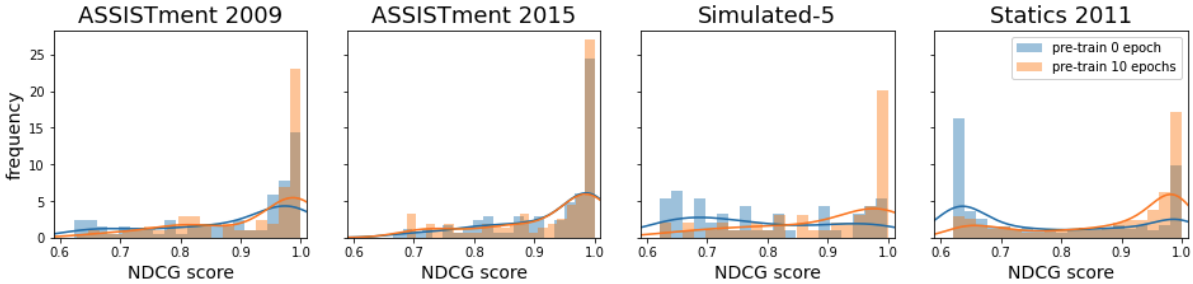


Figure 4. Quantitative inverted prediction (hard) effect with NDCG score distributions. Bars represent the histograms and solid lines are their kernel distribution estimates (KDE). Horizontal axis shows NDCG scores and vertical axis shows frequencies.

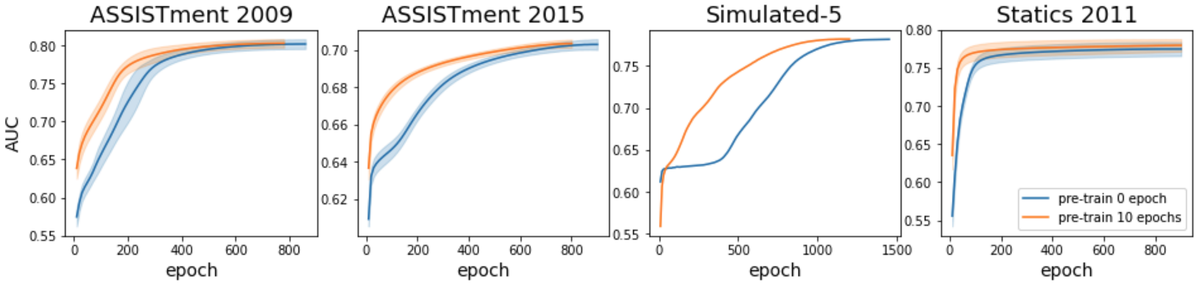


Figure 5. Learning curves of AUC for validation data. Solid lines represent the average of cross-validation. Shaded regions indicate 1-sigma intervals. Horizontal axis indicates the number of epochs and vertical axis the AUC scores.

4.3 Discussion

By pre-training, the model's parameters presumably reached a better starting position before being supplied with real data. The starting position has a property that when a student correctly answers several questions for a specific KC in a row, he/she is likely to answer another question with the same KC correctly. The model may learn this property if the dataset is large. However, the amount of data in educational data mining is often limited, so providing the property as prior knowledge is an effective strategy.

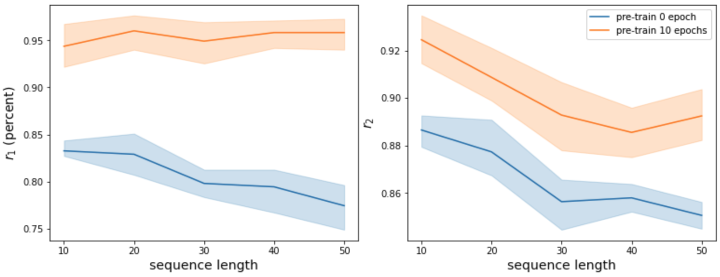


Figure 6. Dependence of r_1 and r_2 on sequence length in ASSISTments 2009. Horizontal axis is the sequence length and vertical axis is the value of r_1 and r_2. Solid lines show the average prediction probability for all KCs. Shaded regions represent $\mu \pm \sigma$ (i.e., mean $\pm$ standard deviation).

The results show that pre-training regularization had a positive effect on the predictive performance as well. With pre-training, the starting AUC for real data training is higher. The model also got a higher maximal validation AUC. For ASSISTments 2009, ASSISTments 2015, and Statics 2011, improvements in AUC scores were 0.2 ~ 0.7 %, which are significant considering the scores are already high (around 70 ~ 80 %). This result suggests that it had a better training starting point, resulting in a more global minimum and a better score. Simulated-5 is a simulated dataset and may not correctly reflect the learning dynamics of real students. Further investigation is needed to explain why the test AUC decreases with pretraining. For Statics 2011, pretraining 10 times produced better test AUC than pretraining 150 times. Such a decrease in test AUC can be explained by prior working too strong. Optimizing the amount of pretraining is a necessary step that we should explore.

We also examined the r_2 score further. We calculated the NDCG score using synthetic sequences for each KC, as described in Section 3. The NDCG scores can be visualized as a distribution. With pre-training, the number that equals or nearly equals 1 has increased. It means that the number of predictions having a perfect order has increased. Also, the distribution shifted toward 1. It suggests that our method made the prediction results better for most KCs.

Figure 3 shows the prediction results for synthetic inputs. It shows improvements for prediction results having counter-intuitive values of r_1 and r_2. The line plot gets smoother as the number of correct answers in the synthetic sequence increases from 0 to 20.

In Figure 4, the result for Simulated-5 indicates an interesting feature: without pre-training regularization, its r_1 score is only about half. It suggests that the dataset may have failed to reproduce the proportional prediction rule and that the DKT model can still perform well without the data coming from real human responses. In contrast, the model's AUC performance on Simulated-5 data dropped when using pre-training regularization.

Figure 5 shows the learning curves. When using pre-training, the model had a higher AUC at the start of the training iteration, and it continued to get a higher AUC than the baseline. It means that the model can get a considerable gain at an early stage. From another viewpoint, the result suggests that the vanilla DKT may be using the first few epochs just to acquire the proportional prediction rule. It would be more efficient to make the DKT model learn the rule using a few epochs of synthetic data since they're much smaller in size and take less time to train. Although our proposed method requires some epochs of pre-training, this is not a massive disadvantage in terms of the training time since it scales with the number of KCs, which is a much smaller number than the size of the training data. Also, the model becomes less likely to overfit, as indicated in the later epochs.

Figure 6 shows how r_1 and r_2 changed with respect to the sequence length T. The model with pre-training for 10 epochs generally had higher r_1 and r_2 than the model without pre-training. As T increased, both r_1 and r_2 decreased. This suggests that the inverted prediction problem becomes more significant for longer sequences.

5. Conclusion

In this paper, we pointed out a new difficulty in DKT: the inverted prediction problem. To alleviate the problem, we proposed pre-training regularization to train the model using synthetic data before providing real data to the model.

The method is easily adapted to any model since it does not require model adjustment or introducing a prior distribution, as done in Bayesian modeling. The experiments showed that pre-training regularization can reduce the effect of inversely proportional predictions quantified using two indicators, r_1 and r_2. It also made the model generalize better and resulted in higher AUC scores.

In the real world, there are various ways of utilizing the predicted academic abilities of students. It is not enough to evaluate the model by its accuracy when teachers or students ask to see the prediction scores. The prediction in such a situation should match the user's intuition, as otherwise, the system cannot gain much trust from practitioners of education. It is crucial to investigate the prediction in detail and check it to human intuition.

Our method's strength is that it can incorporate various types of prior knowledge through the design of simulation algorithms. It is advantageous when we want to use a prior distribution that has no efficient sampling algorithm. It is also easier for education researchers to incorporate their knowledge on how students learn. Researchers only need to provide rules or specific examples and put them into the DKT model. They are not required to have in-depth knowledge of probability theory. For example, they can generate samples reflecting a tendency that a certain KC is likely to be acquired after another KC. They do not need to represent it as a stochastic model. If the prior knowledge is correct, it can increase the DKT model's performance, for example, in higher AUC, as it did in our case with the proportional prediction rule.

One limitation of our approach is that problems must be assigned to properly defined skills. If the assignments are unreliable, the underlining assumption of monotonic increase of the scores cannot be used. In future work, we plan to introduce knowledge state vectors (KS vectors) used in our previous work to represent the levels of skill acquisition (Pan and Tezuka, 2020). Since KS vectors can take continuous values, it can add more flexibility to the model than representing skill acquisition by binary values, as we did in this work. We also consider comparing how the order of supplying synthetic and real data may affect the performance. Preliminary experiments showed that supplying synthetic data before real data was more effective than supplying them after real data. We want to investigate this difference further and examine its validity for different types of tasks and models.

Acknowledgements

This work was supported by G-7 Scholarship Foundation and JSPS KAKENHI Grant Numbers JP16K00228, JP18KK0308.

References

Adamopoulos, P. (2013). What makes a great MOOC? An interdisciplinary analysis of student retention in online courses. *Proceedings of the 34th International Conference on Information Systems*.

Baker, R. S. J. d., Corbett, A. T., & Aleven V. (2008). More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, 406-415. Springer.

Cen, H., Koedinger, K., & Junker, B. (2006). Learning factors analysis - A general method for cognitive model evaluation and improvement. *Proceedings of the 8th Int’l Conference on Intelligent Tutoring Systems*, 2006.

Corbett, A. T. & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-adapted Interaction*, 4(4), 253-278.

Ding. X. & Larson, E. C. (2019). Why deep knowledge tracing has less depth than anticipated. *Proceedings of the 12th International Conference on Educational Data Mining*.

Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, 19(3), 243-266.

Kaplan, A. M. & Haenlein, M. (2016). Higher education and the digital revolution: About MOOCs, SPOC, social media, and the cookie monster. *Business Horizons*, 59(4), 441-450.

Khajah, M., Lindsey, R. V. & Mozer, M. C. (2016). How deep is knowledge tracing? *Proceedings of the 9th International Conference on Educational Data Mining*.

Pan, Q. & Tezuka, T. (2020). Accuracy-aware deep knowledge tracing with knowledge state vectors and an encoder-decoder architecture. *Proceedings of the 28th International Conference on Computers in Education*.

Pan, S. J. & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345-1359.

Pardos, Z. A. & Heffernan, N. T. (2011). KT-IDEM: Introducing item diffculty to the knowledge tracing model. *Proceedings of the 19th International Conference on User Modeling, Adaptation, and Personalization*, 243-254. Springer.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. *Advances in Neural Information Processing Systems*, 505-513.

Resnick, L. B. & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. *Changing Assessments*, 37-75, Springer.

Ritter, S., Harris, T. K., Nixon, T., Dickison, D., Murray, R. C., & Towle, B. (2009). Reducing the knowledge tracing space. *International Working Group on Educational Data Mining*.

Sapountzi, A., Bhulai, S., Cornelisz, I., van Klaveren, C., Kardaras, D., & Semanjski, I. (2018). Dynamic models for knowledge tracing and prediction of future performance. *Proceedings of the 7th International Conference on Data Analytics*.

Sawade, C., Bickel, S., von Oertzen, T., Scheffer, T. & Landwehr, N. (2013). Active evaluation of ranking functions based on graded relevance. *Proceedings of the 23rd Int'l Joint Conference on Artificial Intelligence*.

Shepard, L. A. (1991). Psychometricians' beliefs about learning. *Educational Researcher*, 20(7), 2-16.

Wang, Y., Wang, L., Li, Y., He, D., & Liu, T.-Y. (2013). A theoretical analysis of NDCG type ranking measures. *Proceedings of the 26th Annual Conference on Learning Theory*, 25-54.

Wilson, K. H., Karklin, Y., Han, B., & Ekanadham, C. (2016). Back to the basics: Bayesian extensions of IRT outperform neural networks for proficiency estimation. *Proceedings of the 9th International Conference on Educational Data Mining*.

Xiong, X., Zhao, S., Van Inwegen, E. G., & Beck, J. E. (2016). Going deeper with deep knowledge tracing. *Proceedings of the 9th International Conference on Educational Data Mining*.

Yang, T.-Y., Brinton, C. G.. Joe-Wong, C., & Chiang, M. (2017). Behavior-based grade prediction for MOOCs via time series neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(5).

Yeung, C.-K. & Yeung, D.-Y. (2018). Addressing two problems in deep knowledge tracing via prediction-consistent regularization. *Proceedings of the 5th Annual ACM Conference on Learning at Scale*, 5, ACM.

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized Bayesian knowledge tracing models. *Proceedings of the 16th International Conference on Artificial Intelligence in Education*.

Zaremba, W., Sutskever, I., & Vinyals, O. (2014). Recurrent neural network regularization, arXiv:1409.2329.

Zhang, L., Xiong, X., Zhao, S., Botelho, A., & Heffernan, N. T. (2017). Incorporating rich features into deep knowledge tracing. *Proceedings of the 4th ACM Conference on Learning at Scale*, 169-172. ACM.

From Hello to Bye-Bye: Churn Prediction in English Language Learning App

Daevesh SINGH*, Rumana PATHAN, Gargi BANERJEE & Ramkumar RAJENDRAN
IDP in Educational Technology, Indian Institute of Technology Bombay, India
*daeveshsingh@iitb.ac.in

Abstract: Mobile phones and apps have changed the landscape of e-learning and have revolutionised the way people learn a second language by facilitating anytime-anywhere learning, game-based resources and socially interactive learning activities. Despite these features and affordances, these language learning apps suffer a fate of high churn rates. In this paper, we examined the churning behaviour of learners in the context of a language learning app called Hello English. We applied descriptive analytics to analyse the behavioural differences between churners and non-churners and studied their interaction with the app to early-predict churning behaviour. Our findings indicate that non-churners interact with the mobile app more frequently compared to churners. Also, the trained machine learning classifiers can predict learner churning behaviour with a high recall value (0.824) and F1 (0.778). This churn detection will enable the app developers to provide intervention for learner retention.

Keywords: Language learning, churn rate, interaction behaviour, descriptive analytics

1. Introduction

The popularity of mobile phones has immensely increased in the last two decades due to various factors like portability, unobtrusiveness, ease of use, affordability, and personal adaptation (Chih & Shih, 2011; Sharples, 2000). The installation of applications (apps) on these devices further enhances their usability and capability and have made them a preferred tool for entertainment, business and learning (Godwin, 2011; Papadakis et al., 2020). These apps, which were initially designed for assisting productivity (emails, calendar etc.), are now being explored in other areas such as gaming, online shopping, social media, and education (Papadakis et al., 2020). The apps developed for educational purposes provide enormous educational resources, interactive activities, challenge-based learning through games, puzzles and collaborative activities, which have changed the landscape of e-learning, especially in the field of Mobile-Assisted Language Learning (MALL) (Burston, 2014; Chih & Shih, 2011; Kim & Kwon, 2012; Miangah & Nezarat, 2012).

MALL has revolutionised the way people learn a second language by facilitating anytime-anywhere learning, collaborative learning activities, socially interactive and game-based resources (Kim & Kwon, 2012). Despite these features and affordances, MALL faces a plethora of challenges, such as mismatch between pedagogy and technology resulting in fragmented language practice (Pareja et al., 2013), focus on receptive language skills denying the opportunities for socio-cognitive activities (Kim & Kwon, 2012), and low motivation leading to high attrition rates (Elaish et al., 2019). These issues point out the need to employ behavioural analytics to understand how learners interact with these language learning apps (LLAs). Analysing the learner behaviour can help us throw light upon the various features of the LLAs that learners are exploring. For example, the time they are devoting to use these LLAs, how they interact with other learners in LLA, and their progress. Such information about the learner can help us in addressing the issues stated above. Although these issues are crucial to address, this study aims to only focus on analysing learner behaviour to explore the churn behaviour of the learners. Analysing churn behaviour is important as research shows that educational apps have the highest user churn rates among other mobile apps (Pham & Wang, 2016).

App users are said to be churned when they become inactive for a certain period or altogether stop using the app by uninstalling it. The method of identifying churners is called churn prediction or churn behaviour prediction. In our study we have defined churn as week long period of inactivity. Churn prediction is crucial as it will help the app developers devise a strategy to bring back the churners. This is important since churning has a detrimental impact on learning, as the learner's engagement with the app for a more extended period is crucial for learners (Kim & Kwon, 2012; Burston, 2014).

In this paper, we explored the interaction behaviour of 700 randomly chosen learners of an English language learning app called Hello English (HE). The main goal of this paper is to:

1. Find the differences in the interaction behaviour of the churners and non-churners.
2. Use this interaction behaviour to train Machine Learning (ML) models to predict churn behaviour.
3. Early predict the churn behaviour using the best model.

To find the differences in the interaction behaviour of the churners and non-churners, we applied descriptive analytics and compared the average activities per day for both the groups. We used ML models and trained these models on the interaction data of 7 days (observation period) to predict the churn behaviour. Finally, we performed early prediction using ML algorithm. We found that 1) non-churners, on average, performed more activities as compared to churners, 2) the ML model (LR) used for prediction had high recall (0.824) and F1 (0.778) value which is comparable to churn prediction models used in other domains, and 3) this model could early predict the churning behaviour of learners with a recall (0.704) and F1 (0.675) value from the fourth day onwards.

In the sections that follow, Section 2 examines the existing literature on churn prediction in mobile apps, helping us operationalise churn in Section 3. Section 4 provides the details of the learning environment. Section 5 and 6 describes the details of the methods used in this study and the results obtained. Finally, in Section 7, we present the conclusion and limitations of the study.

2. Background and Related Work

Existing research on user retention for mobile apps has progressed in three primary directions, they are:

1. Why learners uninstall apps
2. Finding solutions to enhance user retention
3. interpret the usage behaviour of churners

In the first direction, research studies have found a host of reasons that contribute to the low user retention for mobile apps. In one such study, Ickin et al. (2017) surveyed users and identified reasons that affected user retention like intrusive advertisements, boredom, lack of updates of the app, high memory allocation, and low popularity in their friend circle.

In the second direction, research focused on finding the solutions to enhance user retention. These studies examined many factors affecting user engagement. They suggested that integrating gamified components in the learning design, experiential learning, incorporating digital leaderboards and badges, built-in social media, and frequent release of updates leads to an increase in engagement (Pechenkina et al., 2017; Pham & Wang, 2016).

In the third direction, research studies are focused on analysing the usage behaviour of churners for making app recommendations (Shang et al., 2017), providing scaffolding to learners (Pishtari et al., 2019), or developing models that predict the churning. Although in academia, behavioural analytics is ubiquitous and is a subject of widespread interest ranging from learning, affect detection, dropout and engagement (Nishane et al., 2021; Rajendran et al., 2013, 2018; Pathan et al., 2019, 2020). We did not find any research study focused on developing models for predicting churn in an LLA to the best of our knowledge. Thus, despite mobile educational apps having the lowest retention rates (Pham & Wang, 2016), there is a paucity of literature that explores behaviour patterns of churners in the context of mobile educational apps. In this paper, we focus on interpreting the usage behaviour of churners in the context of a commercial English language learning app.

Although several LLA’s exist, we did not find any research studies related to churn prediction. Thus, we reviewed the research articles on churn prediction in non-educational apps to understand churn and the different features used for churn prediction. The study of Hadiji et al. (2014) predicted churn in mobile games by analysing data of 50,000 randomly selected players from 5 different mobile games up to a specific date (cutoff date). The authors' defined churn in two ways, i.e. "hard churn" and "soft churn". In the case of hard churn, a player without any session after the cutoff date was termed as the churner, and in the case of soft churn, a player with a low number of sessions was considered churned. The feature set included universal features (game-independent features) such as the number of sessions, days, the time elapsed since the last session, average playtime per session and average time between the sessions. They also incorporated four "economy features", such as the number of purchases. The four ML classifiers employed were LR, Neural Networks (NN), Naive Bayes (NB), and Decision Tree (DT). DT predicted churn with high accuracy for some games. However, a relatively low accuracy was reported in games with maximum players who churned after playing a few sessions.

Similarly, the study by Kim et al. (2017) is based on churn prediction of three mobile and online casual games using log data of 193,443 players. The authors have defined churn using observation period and churn prediction period. The user is considered churned if they remain inactive in the churn prediction period after playing at least one session in the observation period. The models included three traditional ML algorithms LR, Gradient Boost (GB), and Random Forest (RF), along with two Deep learning algorithms, i.e., Long Short Term Memory (LSTM) and Convolutional Neural Networks (CNN). A total of ten universal features like the number of sessions played, the time between first and last session, the time between the consecutive sessions etc., along with game-specific features like the number of purchases made, the price paid etc., were extracted. The results show that LSTM outperformed other models in churn prediction for one game with an AUC (i.e. area under the receiver operating characteristic curve) of 0.792. In another game, gradient boosting performed best with an AUC of 0.728. Likewise, in the third game, LR and GB outperformed others with an AUC of 0.842.

In another study, Runge et al. (2014) predicted churn for high-value players of two social games with a data set of 18445 players. In this study, the players are said to churn if they remain inactive for more than 14 days. The authors categorised the data into three types, i.e. in-game data (log data of the player during the play session), revenue data (revenue generated by the player), and player profile data (profile data of the player). The feature set includes features like days in-game, last purchase, days since last purchase etc. The binary classification of churners and non-churners was done using four ML algorithms, viz. NN, SVM, NB, and LR. NN outperformed other models in both games with an AUC value of 0.815 in one game and 0.930 in another.

To summarise, the research papers presented above are based on churn prediction of non-educational app users by analysing the interaction data generated over a period of time. The way churn is operationalised in these studies vary due to the difference in the context. Similarly, the feature sets used in these studies are universal (such as the number of sessions) and app-specific (such as the frequency of using a particular resource). Likewise, the best performing ML models are different in each of these studies. Hence, defining churn and developing a feature set in the context of LLA is a significant research gap that this paper addresses. Since there is no single ML model that outperforms other models in the literature, we are motivated to explore multiple ML models.

3. Defining Churn

There are several ways in which churn is operationalised in the literature, the simplest one being the app's uninstallation. However, this way of defining churn seems inappropriate in the educational context since existing studies have identified the positive impacts of apps on language learning (Burston, 2014; Kim & Kwon, 2012). Therefore, the learning gain of the learners is dependent on the time spent on the app. Hence, being inactive on the app is equivalent to uninstalling the app. As a result, we define churn as a specific period of inactivity.

We did not find any particular method in the literature informing us about choosing an optimal churn period. Also, there is a lack of uniformity in the duration of the churn period reported in the literature. Hence, we define churn as a week-long period of inactivity because our target audience is school-going children (i.e. 14-18 years). Thus, a period of one week will ensure that we monitor them on both working days and weekends. This definition also enabled us to stratify the learners into two groups, i.e. churners and non-churners.

The churn prediction problem requires the interaction data of the learner with the app. As a result, we use seven days of interaction data of the learner, starting from the first day the learner interacts with the app, followed by seven calendar days. This seven day period is called the observation period. For instance, if a learner installs the app on 01 August and interacts with the app on the same day, the period from 01-07 August is considered an "observation period". The seven day period after the observation period (i.e. 08-14 August) is termed the churn period. If a learner does not interact with the app in this period, they will be called churners. A non-churner is a learner who interacts at least once with the app during the churn period.

4. Learning Environment: Hello English

Hello English (HE) is an English learning application specifically designed for second language learners in India to learn English. The application, launched in October 2014, encompasses learning activities that are interactive, personalised, and contextualised to local learning. It encapsulates all four aspects of language acquisition: reading, writing, listening and speaking. The learning activities in HE are equipped with advanced voice recognition technology that allows learners to interact with the app and hold real-life conversations. The interactive learning activities consist of games (individual and multiplayer), activities to practise speaking and contextual learning tools that leverage news, sports, and entertainment. Moreover, most of the app's features can be accessed offline, which saves data expenses for learners and helps make learning a seamless experience.

As mentioned above, learners can perform several activities in HE, which can be broadly grouped under eight categories. That is, learn a lesson (L), practise reading (PR), practice speech (PS), play a game (PG), play reward activity (RA), take a test (TT), respond to a quiz (RQ), and seek help (SH). The app also captures learner data, such as the number of coins for each activity. The detailed description of each category of action is narrated in Table 1.

Table 1. *Categorising Learner's Interaction in Hello English Mobile App and their Description*

	Category of actions	Description
1	Learn (L)	Learner accesses a lesson that involves all the four skills of language learning, i.e. listening, speaking, writing and reading
2	Practice reading (PR)	The learner reads various contextual articles such as current news, articles on entertainment, sports, etc.
3	Practice speech (PS)	The learner practices conversation using inbuilt voice recognition technology and holds real-life conversations.
4	Play a game (PG)	The learner plays various games such as rearranging jumbled words, games that reinforce learning from the preceding lesson etc.
5	Reward activity (RA)	The learner earns coins as rewards for solving different activities.
6	Take a test (TT)	The learner takes a summative assessment after learning a certain number of lessons. These are mainly grammar and vocabulary tests.
7	Respond to a quiz (RQ)	Learner responds to quiz
8	Seek help (SH)	The learner wants to connect with an educator or wants to share his achievement score on social media
9	Number of Coins (C)	It refers to the coins earned after the successful completion of an activity.

5. Research Methodology

The research goal of this paper is to:

1. Identify the difference in the interaction behaviour of churners and non-churners.
2. Use the learner's interaction behaviour to predict whether the learner will churn or not.

3. Reduce the observation period up to 4 days to early-predict churn.

In this section, we first describe the dataset and provide information about the data collection process. The following subsection gives details about data pre-processing procedures used in our analysis. The last subsection informs about the labelling of data.

5.1 Dataset

The dataset analysed in this paper for churn prediction is obtained from randomly selected 700 learners of HE. These learners installed the application between July 2019 and September 2019 and were aged between 14 to 18. At the time of app installation, the learner's consent was sought, and they were informed about all the terms and conditions they need to agree to for using the app. We collected the data comprising all the activities (as described in table 1) completed by the learners during their observation period.

5.2 Data Pre-processing and Analysis

The data consisted of 700 learners who interacted with the language learning application for a minimum of two sessions. The number of days for which they interacted with the app is different for different learners, varying from 1 to 90 days. Out of these 700 learners, more than 90 per cent of learners had more than ten sessions. In the data-preprocessing stage, we discarded the possibly erroneous data (e.g. sessions that lasted for days). We then normalised the data by rescaling the features to the range between 0 and 1 using max-min normalisation. This normalisation is crucial as the data had varying scales. Finally, we used this processed data for extracting features.

To perform descriptive and predictive analytics, labelling learners was crucial. So, we provided labels to these learners as per the definition of the churn described in section 3. The churners were learners with zero sessions in the churn period. Out of the 700 learners, 347 were churners, and the remaining 353 were non-churners.

6. Results and Discussion

This section presents the result of the study by 1) providing a comparison of the interaction behaviour of churners and non-churners, 2) describing features extracted from the interaction data and performance of ML models employed to predict learner churn behaviour, and finally, 3) reporting the best performing model to early to predict learner churn behaviour.

6.1 Comparing Interaction Behaviour of Churners and Non-Churners

In order to compare the interaction behaviour of churners and non-churners, we computed the average number of actions per day per learner in each category of activities. For instance, we calculated the total number of 'practice game' activities per day per learner by dividing the total number of practice games by the product of the number of active days and the number of learners in that group. By ‘active days’, we mean days on which learners perform at least one activity. Figure 1 represents the frequency distribution of the average actions per day per learner for churners and non-churners. The frequencies represent the average number of actions per day per learner in each group. On average, churner has completed only 7.04(SD=3.5) number of activities per day compared to non-churners who have completed 9.80(SD=5.9) number of activities per day. Similarly, churners have accessed only a 3.23(SD=2.68) number of lessons per day on average as compared to 4.27 (SD=4.7) number of lessons by non-churners. We obtained similar results for practice games as well.

Overall, non-churners have accessed more activities per day on average than churners in all the action categories. Our following subsection used these differences in interaction behaviour to predict the learner's app uninstallation behaviour.

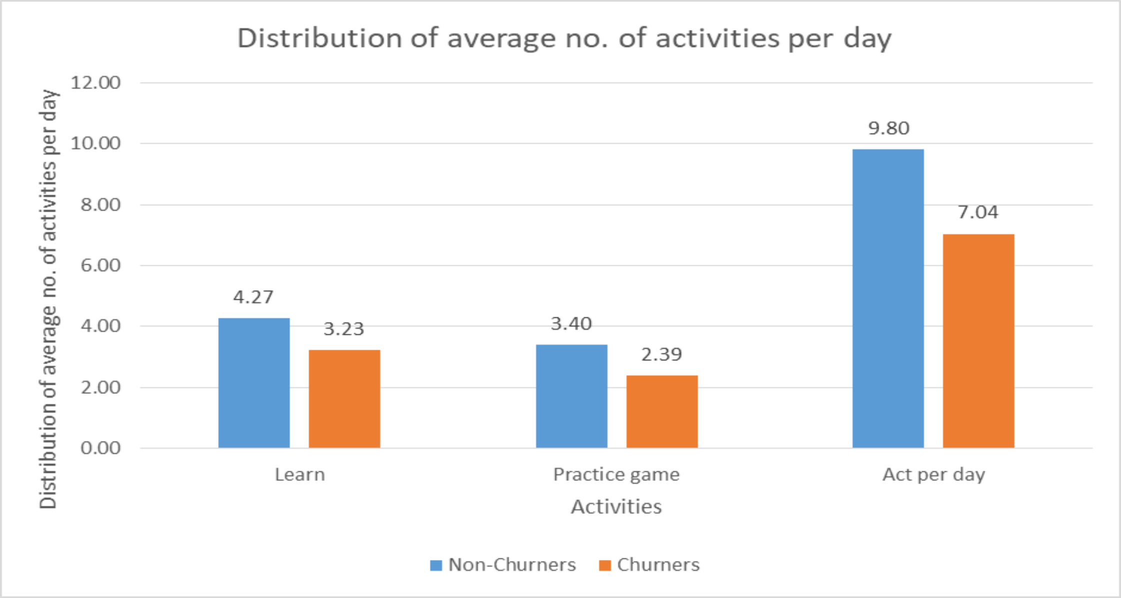


Figure 1. Frequency Distribution of per day per User Activities for "Churners" and "Non-churners"

6.2 Predicting Learner's Churn Behaviour

To predict the churn behaviour, we computed the features using the actions categorised in table 2. Table 2 describes all the features developed from the log interaction data of learners during the observation period.

Table 2. *List of Features Extracted from Interaction Data and their Description*

	Category of actions	Description
1	Active_days	Days in the observation period on which the learner did at least one activity.
2	Inactive_days	Days in the observation period on which the learner did not interact with the app.
3	Lesson/day	The number of lessons learned is divided by the number of active days.
4	PR/day	The number of reading lessons practised divided by the number of active days.
5	PS/day	The number of speech lessons practised is divided by the number of active days.
6	PG/day	The number of games played is divided by the number of active days.
7	RQ/day	The number of responses to the quiz divided by the number of active days.
8	RA/day	Number of reward activities divided by number of active days.
9	C/day	Number of coins earned divided by number of active days
10	A/day	Total number of activities divided by the number of active days.
11	C/activity	Number of coins earned per activity.

We developed these features along two dimensions, namely the learner activity dimension and the reward activity dimension. For example, in the learner activity dimension, we extracted features such as number of lessons learned per day (L/day), number of reading lessons practised per day (PR/day), number of speech lessons practised per day (PS/day), number of games played per day (PG/day), number of responses to quiz per day (RQ/day), and the total number of activities per day.

Similarly, we considered features such as reward activity per day (RA/day), number of coins per day (C/day), and the number of coins per activity (C/activity) as features emerging from the reward dimension. We extracted these features for both groups and developed a classifier using three major classification algorithms proven efficient for small datasets (Sharma & Paliwal, 2015). The classification algorithms used in this research are RF, NB, and LR.

Table 3 shows the result of the model's prediction using 10-fold cross-validation on three different classifiers. We stratified the data at the student level. The results indicate that the LR algorithm performed better than other algorithms in terms of all the evaluative metrics, i.e. F1 score (0.778), precision (0.800), recall (0.824) and accuracy (0.824).

Table 3. *Performance of Different ML Models in Predicting Churners using the Hello English App Users' 7-Days Interaction Data*

Models	F1	Precision	Recall	Accuracy
Random Forest	0.746	0.730	0.777	0.801
Naïve Bayes	0.740	0.772	0.720	0.720
Logistic Regression	0.778	0.800	0.824	0.824

6.3 Early Prediction of Learners Churn Behaviour

The LR model outperformed the other models in the churn prediction task using seven day observation period. We, therefore, used the LR algorithm for early prediction. This model also used 10-fold cross-validation and student-level stratified data. The performance of the LR algorithm using learners' interaction data with different duration of the observation period ranging from day 4 to day 6.

Since we are making an early prediction of churning, it is crucial to identify the set of learners who have churned. Recall as a measure signifies the proportion of all churners that the model accurately predicted. Therefore, we emphasise recall score because it is preferable not to miss any user about churn, even if the model flags some non-churners as churners. Similarly, the F-score is used to measure the model's performance by considering both precision and recall. It is the harmonic mean of precision and recall. Hence, we plotted the recall and F-score values for different days, i.e. days 4-6 (refer to figure 2 and table 4), to understand how early we can make a reasonable prediction of the churning, comparable to other works in different domains. We did not further reduce the observation beyond this as then the learner interaction data would reduce, which might make the model pick up noise.

Table 4. *The Performance of the Logistic Regression Model using Learners' Interaction Data with Different Duration of the Observation Period Ranging from Day 4 to Day 6*

Days	F1	Precision	Recall	Accuracy
4	0.675	0.685	0.704	0.704
5	0.726	0.750	0.809	0.808
6	0.753	0.773	0.813	0.812

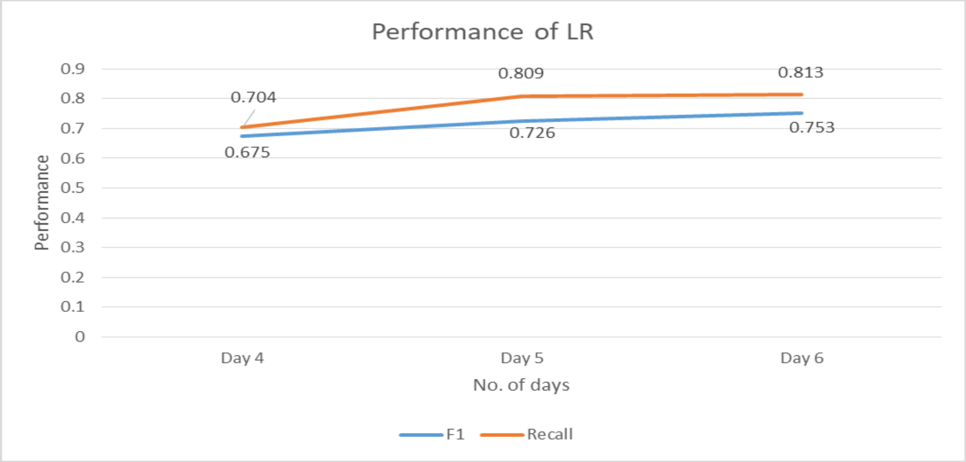


Figure 2. The Performance (F1 Score and Recall) of the Logistic Regression Model using Learners' Interaction Data with Different Duration of the Observation Period (Day 4 To Day 6).

We observed that from day four onwards, we could predict the churning behaviour of learners with a high recall value (0.704) and F1 = 0.675. This result indicates, with data from the first four days, we can predict the learners who will churn and develop focused activities to persist the learners to continue using the app.

7. Conclusion and Limitations

This article is an attempt to make use of learner interaction data for predicting churn behaviour. The significance of our work lies in the fact that it is the first time an educational app is used for churn prediction. Although the study explores the interaction behaviour of learners in a specific LLA called "Hello English", many of the features extracted are generic (features that are not app-specific such as the number of sessions, number of days the app is used etc.). They hence could be easily extended to other LLAs. We analysed the interaction behaviour of churners and non-churners at three levels of granularity:

1. comparing frequencies of various actions performed by them
2. predicting learners churning behaviour using interaction data of 7 days
3. early prediction of churning using different observation periods starting with four days periods.

We found that non-churners accessed a higher number of activities as compared to churners. We could also predict learners churning behaviour using the LR classification algorithm with a reasonably good recall and F-score. Similarly, we have also shown that early prediction is possible with a four-day observation period, making our work significant. Early detection of these learners allows mobile app developers to target a specific learner group and implement engagement measures to retain the learner group.

Although early prediction allows us to predict learner behaviour decently, this approach does not inform us why the learners uninstalled the app. Our future goal is to understand the churners better by identifying why learners uninstall or become inactive for a longer duration. We would like to perform qualitative studies with focussed interviews to identify the reasons for uninstallation and provide informed decisions to the learners. Also, the procedure described in this paper can be further enhanced by using other classification algorithms such as deep neural networks with more data from learners to improve the models' performance further.

Acknowledgements

We would like to thank CSF for funding this project and also Hello English Team for providing data.

References

Burston, J. (2014). MALL: The pedagogical challenges. *Computer Assisted Language Learning*, *27*(4), 344-357.

Chih-Ming, C., & Shih-Hsun, H. (2011). Personalised intelligent mobile learning system for supporting effective english learning [J]. *Educational Technology & Society*, (3), 153-180.

Elaish, M. M., Shuib, L., Ghani, N. A., & Yadegaridehkordi, E. (2019). Mobile English language learning (MELL): A literature review. *Educational Review*, *71*(2), 257-276.

Godwin-Jones, R. (2011). Mobile apps for language learning. *Language Learning & Technology*, *15*(2), 2-11.

Hadiji, F., Sifa, R., Drachen, A., Thurau, C., Kersting, K., & Bauckhage, C. (2014, August). Predicting player churn in the wild. In *2014 IEEE Conference on Computational Intelligence and Games* (pp. 1-8). Ieee.

Kim, H., & Kwon, Y. (2012). Exploring smartphone applications for effective mobile-assisted language learning. *Multimedia-Assisted Language Learning*, *15*(1), 31-57.

Kim, S., Choi, D., Lee, E., & Rhee, W. (2017). Churn prediction of mobile and online casual games using play log data. *PloS one*, *12*(7), e0180735.

Miangah, T. M., & Nezarat, A. (2012). Mobile-assisted language learning. *International Journal of Distributed and Parallel Systems*, *3*(1), 309.

Nishane, I., Sabanwar, V., Lakshmi, T.G., Singh, D., Rajendran, R., Learning about learners: Understanding learner behaviour in software conceptual design TELE., To be appeared in the Proceedings of *International Conference on Advanced Learning Technologies (ICALT), 2021*

Papadakis, S., Vaiopoulou, J., Kalogiannakis, M., & Stamovlasis, D. (2020). Developing and exploring an evaluation tool for educational apps (ETEA) targeting kindergarten children. Sustainability, 12(10), 4201.

Pareja-Lora, A., Arus, J., Mart ın Monje, E., Read, T., Pomposo-Yanes, L., Rodrıguez-Arancon, P., Barcena, E. (2013). Toward mobile assisted language learning apps for professionals that integrate learning into the daily routine. In L. Bradley & S. Thou€esny (Eds.), 20 years of EUROCALL: Learning from the past, looking to the future. Proceedings of EUROCALL 2013 EUROCALL Conference, Evora, Portugal (pp. 206–210). Dublin: Research Publishing

Pathan, R., Rajendran, R., & Murthy, S. (2020). Mechanism to capture learner’s interaction in VR-based learning environment: design and application. Smart Learning Environments, 7(1), 1-15.

Pathan, R., Shaikh, U., & Rajendran, R. (2019, December). Capturing learner interaction in computer-based learning environment: design and application. In 2019 IEEE Tenth International Conference on Technology for Education (T4E) (pp. 146-153). IEEE.

Pechenkina, E., Laurence, D., Oates, G., Eldridge, D., & Hunter, D. (2017). Using a gamified mobile app to increase student engagement, retention and academic achievement. *International Journal of Educational Technology in Higher Education*, *14*(1), 1-12.

Pham, P., & Wang, J. (2016, October). Adaptive review for mobile MOOC learning via implicit physiological signal sensing. In *Proceedings of the 18th ACM international conference on multimodal interaction* (pp. 37-44).

Pishtari, G., Rodríguez-Triana, M. J., Sarmiento-Márquez, E. M., Terasmaa, J., Kori, K., Kangur, M., ... & Puusepp, L. (2019, September). An overview of learning design and analytics in mobile and ubiquitous learning. In *International Conference on Web-Based Learning* (pp. 312-319). Springer, Cham.

Rajendran, R., Iyer, S., Murthy, S., Wilson, C., & Sheard, J. (2013). A theory-driven approach to predict frustration in an ITS. *IEEE transactions on learning technologies*, *6*(4), 378-388.

Rajendran, R., Munshi, A., Emara, M., & Biswas, G. (2018). A temporal model of learner behaviors in OELEs using process mining. In *Proceedings of ICCE* (pp. 276-285).

Runge, J., Gao, P., Garcin, F., & Faltings, B. (2014, August). Churn prediction for high-value players in casual social games. In *2014 IEEE Conference on Computational Intelligence and Games* (pp. 1-8). IEEE.

S. Ickin, K. Petersen, and J. Gonzalez-Huerta, "Why do users install and delete apps? a survey study," in International Conference of Software Business. Springer, 2017, pp. 186–191.

Shang, J., Wang, J., Liu, G., Wu, H., Zhou, S., & Feng, Y. (2017, November). App Uninstalls Prediction: A Machine Learning and Time Series Mining Approach. In *International Conference on Neural Information Processing* (pp. 514-522). Springer, Cham.

Sharma, A., & Paliwal, K. K. (2015). Linear discriminant analysis for the small sample size problem: An overview. International Journal of Machine Learning and Cybernetics.Sharma, A. and Paliwal, K. K., "Linear discriminant analysis for the small sample size problem: An overview," International Journal of Machine Learning and Cybernetics, 2015.

Sharples, M. (2000). The design of personal mobile technologies for lifelong learning. *Computers & education*, *34*(3-4), 177-193.

A Thematic Summarization Dashboard for Navigating Student Reflections at Scale

Yuya ASANO[a*], Sreecharan SANKARANARAYANAN[b], Majd SAKR[b] & Christopher BOGART[b]

[a]*Intelligent Systems Program, University of Pittsburgh, USA*
[b]*School of Computer Science, Carnegie Mellon University, USA*
*yua17@pitt.edu

Abstract: Instructors often ask students to reflect on projects or tasks because it has been shown to be effective for learning. Instructors also use these reflections to improve future offerings of a course. Sifting through reflections manually, however, is both time-consuming and inefficient, especially for large courses. This paper describes a method for organizing student reflections by named entities (i.e., topics of interest) and instructor-defined "themes" to produce summaries that better meet the needs of instructors. Named entities are first extracted from the reflection corpus. Upon choosing one named entity to explore, sentences mentioning that entity are collated from across student reflections. The selected sentences are then classified into instructor-defined themes. Instructors can choose to re-define themes as necessary with support from the system in the form of prevalence statistics and theme-definition suggestions. Finally, a summary of student reflections for each theme is provided. This process and the resulting summaries were evaluated in a semi-structured Wizard of Oz interview study with the teaching assistants of a 160-student graduate-level course on Cloud Computing offered online to the students at Carnegie Mellon University. Results from quantitative Likert-scale analyses and qualitative coding show that teaching assistants preferred our topic and theme-focused summaries over general summaries generated from a random subset of student reflections. Deployment in the form of an instructor-facing dashboard and improvement to the system to allow for uncommonly expressed content to be better discoverable through the dashboard are planned for future work.

Keywords: Instructor dashboard, student reflections, natural language processing, summarization, semi-structured interview, qualitative coding

1. Introduction

Written student reflections either about all or part of a course are a common way of enhancing student learning (Baird et al., 1991; Lee, & Hutchison, 1998; Menekse et al., 2011). For instructors, however, these reflections are rich resources that help them understand what students liked/disliked or found easy/difficult about aspects of the course and improve future iterations of the course. Even though this practice is relatively common, technology-supported ways of sifting through large amounts of unstructured student data have not been effectively addressed in prior work. Manually sifting through this feedback is not just time-consuming but can be subject to instructor bias (Mosteller, 1989). A data-driven way to efficiently navigate written reflections of students, therefore, is an important problem to address.

In order to address this problem, we prototyped an instructor-facing dashboard that provides summaries of student reflections organized by named entities (i.e., topics of interest) and instructor-defined "themes." With named entity recognition and theme labeling, we try to mimic the widely used process of developing codes and sorting them into categories to analyze qualitative data (Erlingsson, & Brysiewicz, 2017). The dashboard first sifts through student reflections to identify the most prevalent named entities (Ex: "Java," "MongoDB," "Scala," and "Spark"). Instructors can choose a named entity to explore, resulting in sentences mentioning this named entity being collated from across student reflections. These sentences are displayed on the dashboard, along with summaries organized by user-defined themes and prevalence statistics about the percentage of student reflections classified into the theme. The entire process can be repeated with different named entities as well as instructor-defined themes. The process, as well as the resulting summaries, were evaluated using semi-structured interviews with the teaching assistants (TAs) of a large online graduate-level course on Cloud Computing[1] offered to 160 students at Carnegie Mellon University. Each TA was responsible for a project unit that students wrote reflections on. They were asked to evaluate the process of defining their own themes and the summaries produced from them. Analysis of Likert-scale questions and qualitative coding of interview transcripts show that the summaries generated by our system are more helpful for instructors in understanding the process students followed when doing assignments than summaries of randomly sampled reflections. We discuss the design implications for summarization tools that better satisfy the needs of instructors.

2. Related Work

There is a significant body of research showing that eliciting student reflections on assignments and lectures helps consolidate learning and improve outcomes (Baird et al., 1991; Lee & Hutchison, 1998). Following this research, several courses embed reflection exercises students can participate in at regular intervals during the course such as the ones by Menekse et al. (2011) and Fan et al. (2017). These reflections not only improve students' understanding of the subject matter but also provide instructors with feedback on their teaching and students' learning. Instructors use these reflections to understand students' experiences in learning and facilitate changes in those experiences (Baird et al., 1991). However, manually coding and summarizing raw written reflections are too laborious for instructors (Mosteller, 1989), and they rarely receive enough support from their institutions to maintain the cycle of analyzing reflections and taking action (Harvey, 2003). Therefore, efficiently parsing and analyzing the content of these student reflections becomes an important problem to solve.

Prior attempts at doing this have been seen, for example, in the work of Fan et al. (2017)'s CourseMIRROR app, where significant work was done not only in exploring the educational benefits of reflection but in extracting insights from the resulting corpus of text. Their app summarizes the reflections they collect not only for instructors to use for course improvement but for students to think about the lecture from multiple perspectives. Their summaries are lists of semantically clustered phrases representing answers to a question about what concepts were "confusing or needed more detail" in a lecture.

Outside of the context of education, there are various algorithms to generate summaries automatically, too. There are two types of summaries: extractive summary and abstractive summary, but we focus on the latter because student reflections are inherently diverse. In the abstract summarization task, neural models have been shown to outperform others (Rush et al., 2015). Even though these models initially targeted relatively short text, researchers have proposed models that can handle a large amount of text such as scientific papers (Beltagy et al., 2020; Cohan et al., 2018; Zaheer et al., 2020).

Their techniques, however, do not attempt to summarize broader themes embedded in responses to more open-ended questions. Yao et al. (2017) give an overview of summarization techniques, saying that most draw on three components: sentence scoring for importance, sentence selection (for coherence, redundancy, and length of final summary), and sentence reformulation (modifying selected sentences into a coherent summary). This technique, we believe, can be used to build a tool to help instructors more efficiently sift through a large number of student reflections. Our work not only presents a first cut summarization of student reflections by theme but also provides instructors with control over choosing topics and their own themes to explore the student reflection data better while enjoying recent advances in neural models of summarization. Our approach utilizes named entity recognition to automatically extract topics of interest from a student reflection corpus and word embedding to help instructors improve their own themes.

[1] http://www.cs.cmu.edu/~msakr/15619-s21/

3. Course Context and Data

This study was conducted in a completely online semester-long graduate-level course on Cloud Computing offered to the students at Carnegie Mellon University and its campuses in Pittsburgh, Silicon Valley, and Rwanda. As a project-based course, a major component of the class is the completion of 10-12 programming projects of significant complexity, using libraries, resources, languages, and tools provided through commercial cloud providers. Students submit solutions to an auto-grading service as many times as they like before the deadline. The auto grader evaluates source code properties, behavior, and performance of students' code and provides feedback allowing students to iterate and improve their solutions.

The course provides significant scaffolding explanations and videos with the projects, separate single-topic primers demonstrating the deployment and use of required technologies, an online textbook teaching underlying concepts, quizzes, small-group activities, and a reflection and discussion forum.

This study focuses on the analysis of the reflection/discussion forum. After each project deadline, students are required to post a reflection paragraph, prompted by the following question:

> *Consider the following topics when creating your post, however, you should never share any code snippets in your reflection:*
>
> - *Describe your approach to solving each task in this project. Explain alternative approaches that you decided not to take and why.*
> - *Describe any interesting problems that you had overcome while completing this project.*
> - *If you were going to do the project over again, how would you do it differently, and why?*
>
> *After completing this task, confirm that your Reflection Score has been automatically updated on the scoreboard before the project deadline.*

The forum's primary intent is to spark reflection and self-explanation. However, it has also been useful as a way of gathering feedback about the project for iterative improvement of the curriculum from semester to semester.

Students are then asked to reply to three other students' reflections. The reflections are only a small part of the project grade, and points are assigned if the student writes anything at all. However, perhaps because the reflections are seen and discussed by other students, students typically reflect substantially on the project.

A large team of TAs helps operate the course; one TA is assigned responsibility for deploying, supporting, and evaluating each project during the term. After the project has been completed and fully graded, the responsible TA presents to the TA group and instructor an overall evaluation of the project, including a summary of students' responses to a post-project survey, student reflections and discussions, and TAs' experiences with students seeking help in the office hours. When analyzing the written student reflections, TAs read all reflections and select a few representative ones to present to the group. They are asked to identify issues raised through the reflections that should be addressed for future offerings of the project.

4. Methods

The dashboard we have designed helps instructors navigate from an overview to thematic summaries and fine details to understand students' thoughts and opinions on the course and improve it. An instructor would use the tool to first investigate at a high-level what topics (such as tools, services, or languages) were most discussed by students and then drill down into each topic to see several thematic summaries of what students said about it, such as difficulty or usefulness. Since instructors may be interested in different aspects of students' perspectives, depending on the instructor's knowledge and concerns about the project, we allow them to create and modify their own themes. Finally, instructors may drill down further to see samples of text unclassified into any themes or browse the raw reflections directly. By investigating topics broadly, then diving into details, instructors can "take the temperature" of the class's opinions of the project, topic by topic, and dive into the details to understand the reasons so that they can make informed recommendations for future improvement.

In this paper, we focus on describing and evaluating the thematic summarization aspect of the dashboard: a mixed-initiative machine learning algorithm for automatically organizing and summarizing student reflections. We perform a two-level classification of each sentence by named entities (section 4.1) and user-defined "themes" (section 4.2). Summaries are then produced for each of these fine-grained categories by the summarizer (section 4.3).

4.1 Named Entity Recognition

To identify the topics of interest in student reflections, we performed named entity recognition using the Python package spaCy (Honnibal et al., 2021). According to the Seventh Message Understanding Conference, "Named Entities (NE) were defined as proper names and quantities of interest. Person, organization, and location names were marked as well as dates, times, percentages, and monetary amounts" (Chinchor, 1998), but, in our context, we treat them as names of tools, techniques, and terminologies students learn in class. Therefore, when extracting the named entities, we focused only on five types in OntoNotes 5 (ORG, PRODUCT, EVENT, WORK_OF_ART, and LANGUAGE) (Weischedel et al., 2013). Named entity recognition allows instructors to quickly grasp which tools, techniques, and terminologies students talk about in their reflections, which is one of the crucial steps in qualitative analysis.

After obtaining a list of named entities for a particular class module, we classify each sentence of reflections into the entities. We do not classify entire reflections as a unit because students often talk about completely different topics within one reflection. For example, a student wrote, "*Learned from last several projects, I started relative[ly] e[ar]ly this time. H[B]ase is more expensive than I expected. In this project, I was able to explore the u[sa]ge of Hibernate Application, RDBMS and NoSQL databases. I get to understand we have to choose specific techniques based on the user scenario."* They talked about HBase in the second sentence, but the preceding and following sentences were not directly related to HBase. Thus, only the second sentence should be considered when we summarize reflections about HBase.

Before going to the thematic labeling in section 4.2, an instructor has to select one of the entities extracted from student reflections to explore further. To help them decide which entity to choose, our system shows them a histogram of the number of appearances of the entities. An example of one such histogram is shown in Figure 1. Instructors can repeat the entire process, as necessary, with different entities to explore each time a selection is made.

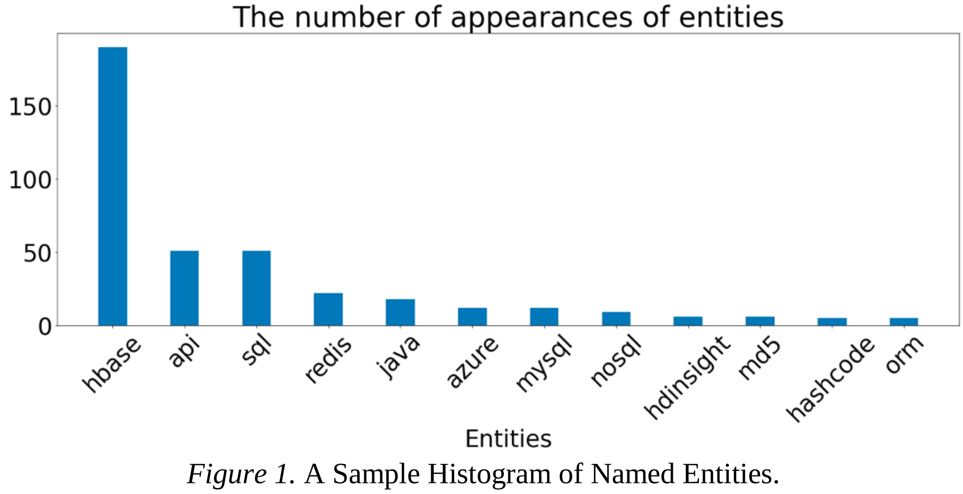
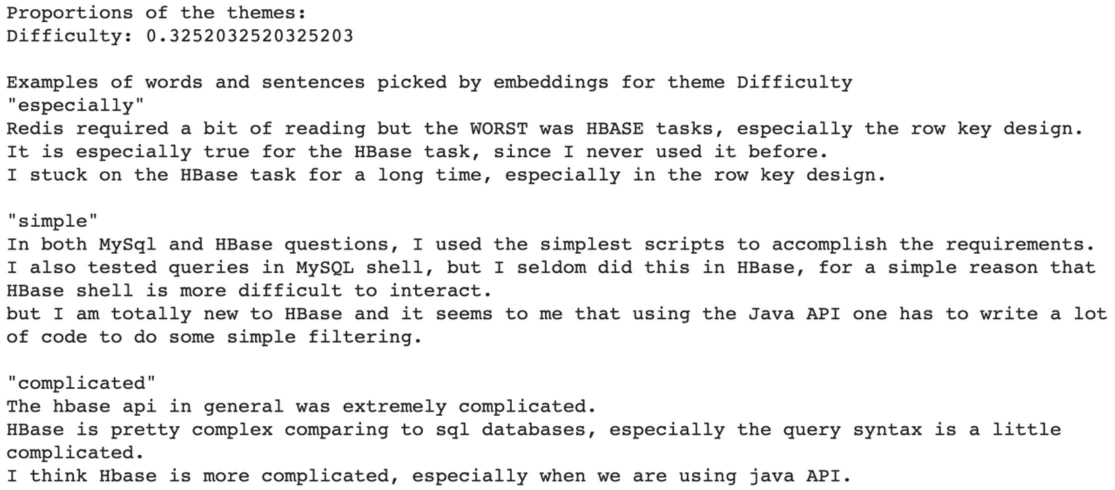


Figure 1. A Sample Histogram of Named Entities.

4.2 Thematic Labeling

After an instructor chooses the named entity that they want to focus on, our system asks them to define a “theme” about the entity. A theme captures how students perceive or feel about the entity selected in the earlier step and is defined by a list of semantically coherent words. For example, the theme “Difficulty” tries to capture if students see the entity as easy or difficult, and it could have the following keywords: “straightforward,” “easy,” “difficult,” “challenging,” and “struggle.” Next, our system classifies the sentences in student reflections that include the keywords into the user-defined theme. For example, if the instructor had chosen HBase in the previous step, the classified sentences would be treated as “difficulty in HBase.” At the same time, the system suggests words in the reflections that have similar meanings to the keywords in the theme, based on the cosine similarities derived from GloVe word embeddings (Pennington & Manning, 2014), by displaying example sentences in reflections that have similar words, as illustrated in Figure 2. Our system also tells the percentage of students who talked about the theme. The instructor may then revise their theme according to the suggestions and the ratio and then run our system again to iteratively improve the theme. In addition, unclassified sentences are also made available to help the TA or instructor later discover other themes related to them.



```
Proportions of the themes:
Difficulty: 0.3252032520325203

Examples of words and sentences picked by embeddings for theme Difficulty
"especially"
Redis required a bit of reading but the WORST was HBASE tasks, especially the row key design.
It is especially true for the HBase task, since I never used it before.
I stuck on the HBase task for a long time, especially in the row key design.

"simple"
In both MySql and HBase questions, I used the simplest scripts to accomplish the requirements.
I also tested queries in MySQL shell, but I seldom did this in HBase, for a simple reason that
HBase shell is more difficult to interact.
but I am totally new to HBase and it seems to me that using the Java API one has to write a lot
of code to do some simple filtering.

"complicated"
The hbase api in general was extremely complicated.
HBase is pretty complex comparing to sql databases, especially the query syntax is a little
complicated.
I think Hbase is more complicated, especially when we are using java API.
```

Figure 2. A Screenshot of Suggestions for Similar Words by Our System. In this example, it suggested adding the words “especially,” “simple,” and “complicated” to the theme “Difficulty.” As a Wizard of Oz study (see Section 5.1), we asked TAs to pick words from the suggestions and manually added to their list of keywords.

4.3 Summarization Technique

We used the Longformer Encoder-Decoder (Beltagy et al., 2020) to summarize the classified reflections. The Longformer was pre-trained to generate abstracts from the papers published in PubMed, using the dataset provided by Cohan et al. (2018). For each named entity and theme, we collected all of the students’ sentences classified into the entity and theme and concatenated them into a single string. This string was the input to Longformer to be summarized.

5. Evaluation

5.1 Participants and Method

We interviewed eleven graduate students who had served as TAs (we call them TA 1 to 11) of the course as described in Section 3. Each TA owned one or two projects in the course in which they had already manually summarized student reflections. We performed a Wizard of Oz walkthrough of a prototype system with each TA, implemented in Jupyter notebook. The Jupyter notebook was operated by the researcher but made visible to the TA over a Zoom video call. In each interview, researchers explained the workflow in section 4 and walked the participant through an interaction with the system, interleaving the requested actions below with explanations and instructions:

- **Choosing a topic**: Participants were first shown a bar graph of the most common named entities found in reflections from their project and were asked to choose an entity to explore further.
- **Choosing theme words**: The researcher then explained the system's concept of *themes* and asked the participant to choose a set of keywords representing a theme.
- **Improving theme words**: The researcher types the theme words into Jupyter notebook python variables, runs the cell, and asks the participant to examine the output (suggested other words that may fit the theme) and revise their set of theme words. Participants iterated this step until they were satisfied with the theme and wanted to go on.
- **Generating a summary**: The researcher triggered the Jupyter cell creating a summary from the final chosen theme.
- **Comparing with a random summary**: The researcher triggers a final cell that shows a Longformer summary of random sentences from reflections, without classification for comparison. We did not use all reflections of a project because there was a maximum number of tokens Longformer could handle at one time (Beltagy et al., 2020).

After TAs interacted with our system and read summaries, we asked the questions in Table 1. Each interview took about 30 to 45 minutes.

Table 1. *Questions asked to TAs during the Evaluation*

Questions	
Q1	(After finishing defining a theme) Did you have any difficulty interacting with our system? Why?
Q2	(After showing our summary) Which parts of the summary are useful to a TA? Why?
Q3	Which parts of the summaries are not useful to a TA? Why?
Q4	How useful would this summary be to a future TA? Rate it on a 7-scale Likert scale (1 is not useful at all, and 7 is very useful).
Q5	What elements did you see in the actual reflections that you wish were included in the summaries?
Q6	(After showing another summary from random reflections) How does it compare to the summary above? Rate it on a 7-scale Likert scale again.

Interviews were performed by one researcher, and most were attended by at least one other researcher. Interviews were transcribed by two researchers and qualitatively coded by a single researcher, identifying 10 themes across the 11 interviews, presented in Table 2; these were discussed and revised with two other researchers who had attended sessions. The most prominent themes are further discussed in the following section.

Table 2. *List of Themes Identified from Interviews with TAs*

Themes	Description
Good summary	Our summary was good.
Unuseful summary	Some parts of our summary were not useful.
Not saying concrete challenges	Our summary did not say concrete challenges students faced.
Something missing	Our summary missed something.
Missing new points of view	Our dashboard may prevent instructors from discovering new perspectives.
Better than random sample	Our summary was better than that of randomly sampled reflections.
Summary of random samples is better	A summary of randomly sampled reflections was better than ours.
Difficulty in thematic labeling	Thematic labeling was easy or difficult.

Intermediate outputs are useful	A list of reflections presented during thematic labeling is useful.
Suggestions for future	TAs made suggestions for new functionalities of our dashboard.

5.2 Results

Overall, TAs told us that our thematic summaries matched their expectations about students' experience. In some cases, they told us our summaries clearly articulated the steps students had followed (Q2). For example, TA 4 stated the evidence that students were able to go through the documentation of syntax and code snippets was useful, highlighting the part of our summary that said Neo4J's documents "are useful and we can quickly pick them up by trying out in the shell." TA 11 told us that the thematic summary showed students' workflow of designing data structure of key-value pairs by using Hadoop map reduction.

When we compared ratings of thematic summaries (Q4, Table 1) with random sample summaries (Q6, Table 1), the average rating of thematic summaries was higher, as shown in Table 3. TAs said that the summaries of randomly sampled reflections did not tell them any new information. For example, TA 8, who rated the thematic summary higher, said the random sample summary discussed Piazza, a Q&A forum used by the course; however, this was course infrastructure, not a topic taught in the class; TA 8 did not find this summary useful because she already knew students relied on it. TA 11, who also preferred the thematic summary, said the random sample summary sounded good but only related opinions and facts the TA already knew:

> *At first glance, [the summary] looks more helpful, but it mostly reinforced assumptions about the students: first time using Hadoop. I already picked that up in office hours. The rest of the pieces are more just summarizing the background about MapReduce*.

Table 3. *The Ratings of our Summaries and Summaries of Randomly Sampled Reflections from Q4 and Q6 of Table 1*

	Our Summaries	Random Summaries
Average Ratings	4.93	3.36
Standard Deviation	1.08	1.85

However, more than half of the TAs thought the summaries of random samples were still useful (Q6) because they included the learning objectives and showed the general consensus of the whole class, and four TAs (TA 2, 5, 6, and 9) rated the random summaries higher than the thematic summaries. For example, one of the learning objectives of the project TA 2 and 6 owned was to differentiate between Spark and MapReduce. Both TAs agreed that the summaries of randomly sampled reflections showed that students had learned this objective, even though instructors often struggled with helping students differentiate between these tools.

In addition, five TAs pointed out that our summaries did not address concrete problems faced by students and the causes of those problems (Q3 and Q4). TA 5 said,

> *"The challenges he faced while implementing the program" [are] something we should solve. ... I'd like to know what challenge that was and if it was something we ignored or we intended to do.*

TA 3, whose rating of the thematic summary was below the average, told us that knowing *why* those problems had happened and *how* students had solved them was important because this could tell TAs what to highlight in the explanations of the projects and where to offer more help to students.

Another concern a few TAs had was that they could only see what they had already expected to see before reading reflections because of the keyword-based theme labeling process. TA 2 worried about failing to catch unexpected things in student reflections. He said he would like to know whether students' struggle was their fault or instructors' fault. TA 4 added that he would be less likely to miss out on anything if he went through every reflection manually. Although our dashboard prototype does save a file of unclassified reflections, TAs did not have a chance to look at these in our study due to the limited interview time.

In terms of the usability of our system (Q1), TAs found it difficult to define themes at first. For instance, TA 9 told us:

> *I was thinking that, for TAs, it's better that you have some hints about what categories could be and for each category what keywords are very likely to be there. I think as a new user to this system, it should take me some time to accommodate. I'd have to try different categories until I realize that I can get some information using such categories but not the others.*

Nevertheless, there is some evidence that this theme definition process can be learned; all three TAs (TA 5, 6, and 7) who had a chance to define two themes rated the second summaries higher or as high as their first try (from 3, 4, and 6 to 5.5, 6, and 6, respectively). Moreover, some TAs stated that they learned useful information during the theme-development process itself; being shown the example sentences and the percentage of the students would be useful because these can reduce the burden of going through unorganized reflections;

> *Basically, every time someone talks about PageRank it's showing here, so I guess that's a good thing. For example, when I'm reading the hundred ... student reviews. You had to do ... that grouping [by] yourself. So, everyone is like 'Okay, these students talk about PageRank.' ... Then you move on to the next and like 'Oh, the students [are] complaining about Scala. Okay,' and that. Having to do that context switch with every student and having to prepare yourself can be tiring* (TA 2).

TA 10 added that it would also be helpful for instructors to know how many reflections are classified into the themes defined by them.

6. Discussion

In this paper, we proposed a novel way to summarize student reflections through two-level classification by named entities and user-defined themes, which is a set of keywords, and compared it with summaries of randomly sampled reflections generated by the same summarizer. Our study with past TAs revealed that our thematic summaries were more useful and better at describing the process students took. In addition, we have found the classification helps TAs reduce their cognitive burden because they can focus on one topic and theme at a time, rather than having to switch contexts constantly if they were to simply read one unrelated student reflection after another. However, this does not suggest that our summary of a selected topic and theme can completely replace summaries of the whole corpus because the latter can show the general consensus of the class. Some TAs indicated such consensus, or even the most commonly used words in all reflections, would be useful when defining themes for our summaries, a task many of them found difficult at first. For example, TA2 said:

> *[It] would be interesting to see the summary [of all reflections] first and then drill down by specific words because I'd be interested to see what's the general consensus with PageRank. ... Imagine, I wasn't expecting complaints about PageRank so [would] be like "oh let's drill down and find those words; why are people complaining, or what they are saying."*

TA 3 suggested showing the most commonly used tokens on the dashboard so that instructors can tell what themes they should define. These testimonies imply that a high-level picture of the entire corpus would help them to make a better selection of keywords in theme labeling (section 4.2).

6.1 Design Implications

Although TAs had a generally positive reaction to the tool, their feedback and our experience building the system suggests several general suggestions for tools such as ours that help build thematic summaries of student reflections:

- **Systems for characterizing student reflections should include thematic summarization** since it appears to help instructors consider topics and themes one at a time rather than context switch between them or conflate issues among them.
- **Thematic summarization should be considered complementary** to tools that allow full browsing and perhaps broad summarization. Instructors need a broad view in order to select

and refine a reasonable theme, as well as to browse and check that they have not built biases into the theme they have selected.

- **Thematic summaries should capture a spectrum of responses**, not a single polar opinion. For example, it appears to be more useful and reliable to summarize students' statements about the spectrum from ease and difficulty together, rather than separately trying to summarize statements that a task is easy from ones that it is difficult.
- **Original reflections behind a summary should be easily accessible** so that instructors can satisfy their curiosity about the reasons and stories behind the statements students make.
- **Unclassifiable responses should also be made visible in some way.** Some TAs said that they sought out unique explanations by individual students that described particular issues, not widely encountered, but worth fixing in course materials. These are easily missed by topic-matching or clustering techniques.

Other possible improvements for future enhancement of the tool include comparing student reflections across multiple semesters as projects evolve and listing the most common negative comments about a tool or service.

6.2 Limitations

There are some limitations to our approach and study. First, our keyword-based method is ambiguous with respect to negation. For example, suppose an instructor defines the theme "Difficult" to gather reflections saying that project tasks associated with a certain entity are difficult. Then, if a student writes the entity "was not difficult," this sentence will be classified into the theme "difficult" even though they mean the opposite. Users can mitigate this effect by including both polarities in their themes. Second, the number of participants (eleven) was too small to conduct statistical testing. This is because we targeted a graduate-level course that typically enrolls fewer students than undergraduate-level courses, hence fewer TAs. Finally, our method can generate only what TAs expect because it asks them to define themes by themselves. Even though our system stores unclassified reflections in a CSV file to help them explore new insights, we could not test its usability because its effective use requires them to define multiple themes. This was not possible in 45-minute-long interviews.

7. Conclusion

Our approach to filtering text for summarization with the interactive entity and keyword selection was considered to be more useful than simply summarizing samples of student reflections in the interviews with TAs and therefore seems to be useful even in its current form. Instructors benefit from our system even without any modifications, but future research can improve it by developing ways to help them pick keywords for themes and discover new themes and tweaking the dashboard to reflect concrete problems faced by students more.

Acknowledgements

We would like to thank TAs who participated in our interviews and students in the cloud computing course who agreed to have their reflections analyzed for this research. This work was funded in part by a grant from Microsoft.

References

Baird, J. R., Fensham, P. J., Gunstone, R. F., & White, R. T. (1991). The importance of reflection in improving science teaching and learning. *Journal of research in Science Teaching*, *28*(2), 163-182.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. *arXiv:2004.05150*.

Chinchor, N. A. (1998). *Overview of muc-7/met-2*. SCIENCE APPLICATIONS INTERNATIONAL CORP SAN DIEGO CA.

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., & Goharian, N. (2018). A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. https://doi.org/10.18653/v1/n18-2097

Erlingsson, C., & Brysiewicz, P. (2017). A hands-on guide to doing content analysis. *African Journal of Emergency Medicine*, *7*(3), 93-99.

Fan, X., Luo, W., Menekse, M., Litman, D., & Wang, J. (2017). Scaling reflection prompts in large classrooms via mobile interfaces and natural language processing. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (pp. 363-374).

Harvey, L. (2003). Student feedback [1]. *Quality in higher education*, *9*(1), 3-20.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2021). spaCy: Industrial-strength Natural Language Processing in Python (Version v3.0.5) [Software]. http://doi.org/10.5281/zenodo.4593273

Lee, A. Y., & Hutchison, L. (1998). Improving learning from examples through reflection. *Journal of Experimental Psychology: Applied, 4*(3), 187–210. https://doi.org/10.1037/1076-898X.4.3.187

Menekse, M., Stump, G., Krause, S., & Chi, M. (2011). The effectiveness of students' daily reflections on learning in engineering context. In *ASEE Annual Conference and Exposition, Conference Proceedings*. https://doi.org/10.18260/1-2--19002

Mosteller, F. (1989). The 'muddiest point in the lecture' as a feedback device. *On Teaching and Learning: The Journal of the Harvard-Danforth Center*, *3*, 10-21.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Weischedel, R., Palmer, M., Marcus, M., Hovy, M., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., El-Bachouti, M., Belvin, R., & Houston, A. (2013). OntoNotes Release 5.0 [Dataset]. *Linguistic Data Consortium.* https://doi.org/10.35111/XMHB-2B84

Yao, J. G., Wan, X., & Xiao, J. (2017). Recent advances in document summarization. *Knowledge and Information Systems*, *53*(2), 297-336. https://doi.org/10.1007/s10115-017-1042-4

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., ... & Ahmed, A. (2020, July). Big Bird: Transformers for Longer Sequences. In *NeurIPS*.

How Can Pedagogical Agents Detect Learner's Stress?

Melanie BLECK, Nguyen-Thinh LE & Niels PINKWART
Humboldt-Universität zu Berlin, Germany
nguyen-thinh.le@hu-berlin.de

Abstract: Learning analytics is aimed to analyze different types of data about the learning process of learners and to provide appropriate pedagogical intervention. However, existing research work on learning analytics mostly focuses on cognitive data (e.g., number of problems solved, number of correct answers, lessons visited) and physical data (e.g., clicks on specific media). The research question to be investigated is how to detect and analyze physiological data of learners in real-time. This paper describes an approach to detecting learners' stress using a pedagogical agent. For that purpose, a wearable wristband sensor is integrated into an existing pedagogical agent for developing human reasoning ability. The pedagogical agent analyzes the heart rate variability of the learner to determine the individual stress threshold. If the learner's stress exceeds the critical threshold, the pedagogical agent offers support with stress coping strategies. The evaluation study with the physiology-aware pedagogical agent shows that the heart rate variability in terms of RMSSD (root mean square of successive differences) can be used as a relevant indicator for measuring learners' stress in real time. The results of the evaluation study suggest deploying RMSSD if stress is taken into account in learning analytics.

Keywords: Learning Analytics, Physiology, Stress, Affect, Pedagogical Agents.

1. Introduction

Learning analytics is "the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (Long & Siemens, 2011). That is, learning analytics is aimed to analyze different types of data about the learning process of learners and to provide appropriate pedagogical intervention. However, existing research work on learning analytics mostly focuses on cognitive data (e.g., number of problems solved, number of correct answers, lessons visited) and physical data (e.g., clicks on specific media) (Chatti et al., 2014; Ferguson, 2013). In addition to the cognitive dimension, other researchers are currently paying attention to the cultural dimension of learning analytics, e.g., (Rüdian et al., 2019). Research on learning analytics using data on the physiological dimension has not gained much attention yet. Several researchers (e.g., Siemens et al., 2011) promoted considering emotional data in addition to cognitive and physical data because the emotional dimension is also important for effective learning. That is, a learner's emotion affects learning motivation and academic achievement (Mega et al., 2014). This paper aims to add learners' physiological data to learning analytics. One of the physiological responses that occur during learning is stress (Li et al., 2017).

The term "stress" has been introduced by Selye who defined stress as "the nonspecific response of the body to any demand for change." (Selye, 1976, p. 15). According to this definition, stress is not emotional valence or nervous tension. Selye differentiated between eustress and distress. The former describes stress that has positive effects and the latter is referred to as stress that is associated with negative effects. In the context of learning, according to this definition, if a learner is in a learning situation whose demand overwhelms her cognitive availability, thus, negative stress or distress would expose. Of course, depending on the individual situation, other negative consequences might also imply such as tiredness, monotony, lowered vigilance, and physical saturation (ISO, 2018). From the physiological point of view, the sympathetic nervous system reacts in a stressful situation and results in physiological responses, for example, the heart rate increases, because specific and relevant muscles will cause stronger blood flow. Academic-related stress can reduce learning achievement. An observational study (Kotter et al., 2017) with 456 medical students has shown that higher perceived academic-related stress was found to predict poor academic performance. Academic-related stress not only has an impact on academic performance but also learners' motivation (Pascoe et al., 2020). Thus, detecting and measuring stress that occurs while learning is relevant for technology-enhanced learning environments.

Different observation techniques have been proposed to track the stress of learners, e.g., facial detection and video monitoring (D'Mello, 2017). Giannacos and colleagues (Giannacos et al., 2020) suggest that physiological parameters, e.g., heart rate, blood pressure, temperature, and electrodermal activity (EDA) level can be used as a proxy to estimate learning performance. The monitoring of physiological parameters like heart rate variability (HRV) is considered a potential indicator for stress detection (Zangroniz et al., 2018). If one is in a stressful state, HRV will decrease (Kim et al., 2018; Mourenas et al., 2018). However, handling physiological data, to what extent they can be used to analyze learners' cognitive demands, and how they can be utilized in a learning context is still a research gap. The research question to be investigated in this paper is how HRV data can be used in real-time by a pedagogical agent to determine the stress level of the learner and to offer the learner stress reduction strategies within the context of technology-enhanced learning environments.

In the next section, the paper presents the methodology for investigating the specified research question. In Section 3, an evaluation study is conducted to test research hypotheses. In the final section, evaluation results are discussed and lessons learned are summarized.

2. Methodology

2.1 A Stress-Sensitive Pedagogical Agent

To investigate the specified research question, the web-based pedagogical agent LIZA (Le & Wartschinski, 2018) that was aimed to improve human reasoning ability, is used. Since the pedagogical agent is intended to measure the physiological states of users, it is required to extend it with three new components. The first component provides a solution to generate, save and process the HRV data. The second one analyzes data regarding stress and the third one adapts the learning situation through selected stress reduction strategies.

2.2 Physiological Parameters of Stress

To use HRV parameters as an indicator for stress, a physiological sensor is required. Such a sensor is provided e.g., by the Empatica E4 wristband that can be considered for a learning environment because it is compact and wearable. These features are required to reduce entry barriers while learning (Gjoreski & Gjoreski, 2017). The integrated photoplethysmography sensor in the wristband is utilized to determine the heart rate and to calculate the time interval between two consecutive heartbeats (i.e., NN intervals) (Empatica Inc., 2016). These data can be transmitted to a server by a mobile application provided by Empatica. The server is responsible for processing HRV and is the first component extended to the existing pedagogical agent LIZA (Le & Wartschinski, 2018).

To determine HRV, Standard Deviation of the NN intervals (SDNN) or the Root Mean Square of Successive Differences of the heart rate (RMSSD) can be applied (in addition to frequency-based parameters such as high frequency or low frequency). SDNN is suitable for long-term measurement that requires a period of about 24 hours, while RMSSD is a short-term measurement that requires between 0.5 and 6 minutes (Sammito, et al., 2014). Thus, RMSSD is chosen as a metric for HRV to be adopted in a pedagogical agent because the learning process through problem-solving can be monitored in short durations of every 5 minutes. RMSSD is calculated using the formula (Figure 1):

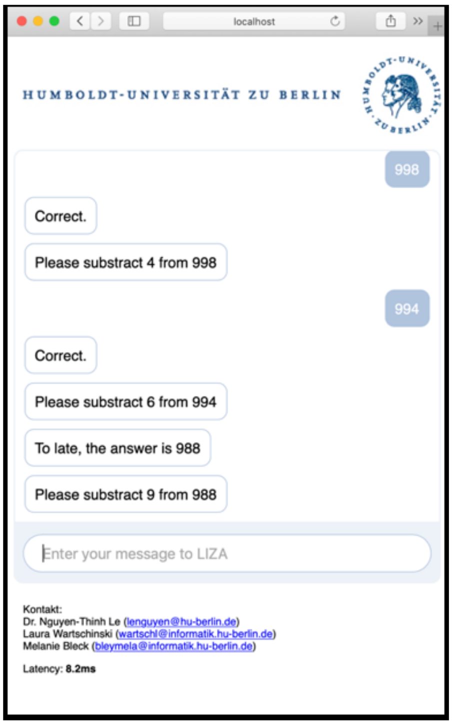
$$\mathrm{RMSSD} = \sqrt{\frac{1}{\mathrm{N}-1}\sum_{\mathrm{i}=1}^{\mathrm{N}-1}(\mathrm{NN}_{\mathrm{i}+1}-\mathrm{NN}_{\mathrm{i}})^2}$$

Figure 1. Root Mean Square of Successive Differences of the heart rate

2.3 Stress Induction

Since there are no generally accepted threshold values to determine a general stress marker for everybody, individual measurements need to be carried out (Sammito, et al., 2014). Thus, a specific individual threshold of stress level needs to be determined. Due to this reason, a stress induction phase was added to the pedagogical intervention process. The second component required to be integrated into the existing pedagogical agent LIZA serves to handle stress induction and stress analysis.

Castaldo and colleagues (2015) conducted a meta-analysis of assessment of stress factors in healthy adults using short-term HRV analysis, i.e., using RMSSD as an appropriate stress indicator. Punita and colleagues (2016) induced stress situations in daily life (e.g., car driving). Examination situations (Tharion et al., 2009) or playing video games (Li et al., 2009) also cause stress. Other stress situations that are more similar to learning settings are the Color-Stroop test (Endukuru et al., 2016; Visnovcova et al., 2014) and arithmetic exercises (Visnovcova et al. 2014; Taelman et al. 2011; Lyu et al. 2015). The stress factors like car driving, examination, and playing video games are not suited to an (online) learning environment that is constructed using a pedagogical agent. On the contrary, Color-Stroop tests and arithmetic exercises can be embedded into (online) learning environments quickly. For the study being investigated in this paper, arithmetic tasks were chosen as an induction method for stress, because they have been used widely to generate moderate stress levels (Schneider et al., 2003).



998

Correct.

Please substract 4 from 998

994

Correct.

Please substract 6 from 994

To late, the answer is 988

Please substract 9 from 988

Kontakt:
Dr. Nguyen-Thinh Le (lenguyen@hu-berlin.de)
Laura Wartschinski (wartschl@informatik.hu-berlin.de)
Melanie Bleck (bleymela@informatik.hu-berlin.de)

Latency: **8.2ms**

Figure. 2. Arithmetic Task in Stress Test 1

The stress induction phase has two arithmetic tests. The first stress test is aimed to derive an RMSSD value for a normal cognitive load. Test 1 represents a normal condition that requires attention for learning. The second stress test, which is designed with a higher difficulty level, is aimed to induce an excessive cognitive overload. Test 2 represents a stress condition. With stress test 1 (see Figure 2), the user has to subtract a random value from a certain number (e.g., starting from 1000) consecutively for five minutes (this period is appropriate for RMSSD measurement). The result of the previous equation provides the minuend of the following. For stress test 1, the time limit for each question is 15 seconds. The pedagogical agent can verify the user's solution as correct or not correct. The second stress test has a similar design but is more difficult to induce a higher cognitive load. The level of difficulty can be altered through three options: 1) reducing the time limit for solving an arithmetic exercise (e.g., less than 15 seconds), 2) increasing the digits of the random subtrahend, and 3) increasing the value of the start minuend (e.g., greater than 1000). The second and third options determine the number of shifts during arithmetic problem solving, which increases the cognitive load with a growing number.

In a range between the normal cognitive demand and the excessive cognitive overload, an individual stress threshold is required to be specified when a user steps into a stressful situation. Considering that an excessive cognitive overload may result in learning demotivation and a drop-out of learning in the long term, the learning situation has to be adapted before such a scenario may occur. Another factor, which has to be taken into account is that an adaption of the learning situation through stress coping strategies will interrupt the learning process itself. So, it should be carried out as little as possible but also as much as necessary.

A preliminary empirical test was conducted in advance, where the task solutions were known and therefore low stress was induced, showed, that a threshold at 50% of the range triggers an intervention nearly every time LIZA was used. This would lead to massive interruptions in the learning process. Based on the results of the preliminary test, the threshold was increased to 2/3 of the individually defined stress range, where the frequency of the intervention could be reduced. The threshold is defined according to the following formula and the stress threshold is illustrated in Figure 3:

$$\text{Threshold} = \max(\text{RMSSD}_{\text{test1}} ;\text{RMSSD}_{\text{test2}}) - (\text{abs}(\text{RMSSD}_{\text{test1}} - \text{RMSSD}_{\text{test2}}) * 2/3)$$

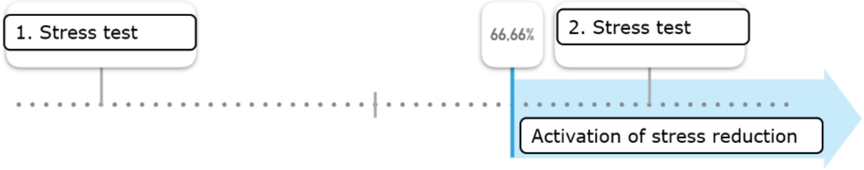
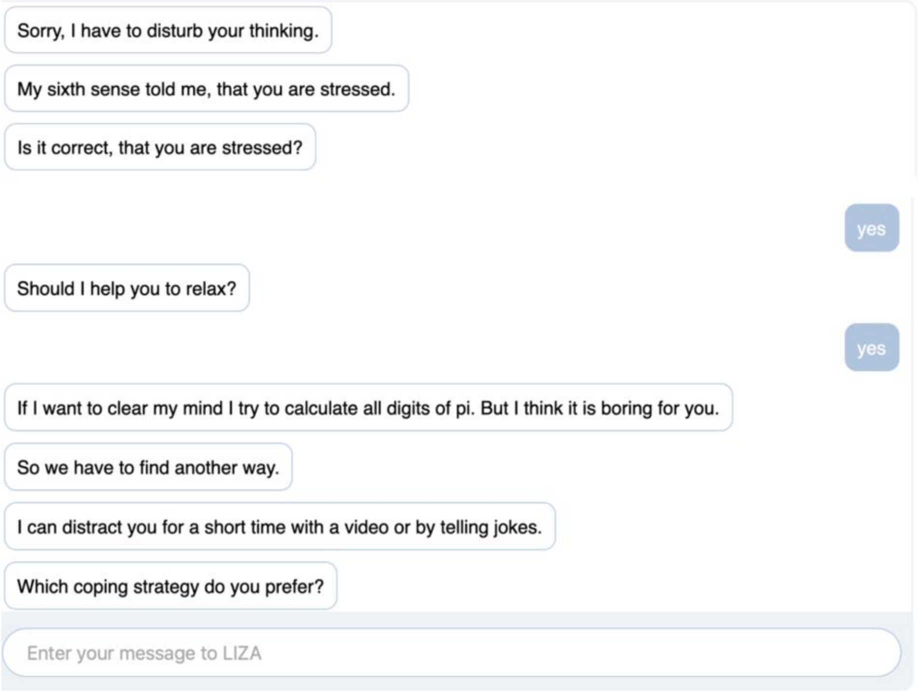


Figure 3. Stress Threshold

2.4 Stress Coping Strategies


Sorry, I have to disturb your thinking.

My sixth sense told me, that you are stressed.

Is it correct, that you are stressed?

yes

Should I help you to relax?

yes

If I want to clear my mind I try to calculate all digits of pi. But I think it is boring for you.

So we have to find another way.

I can distract you for a short time with a video or by telling jokes.

Which coping strategy do you prefer?

Figure 4. The Pedagogical Agent Proposes Two Strategies for Reducing Stress.

If the current RMSSD exceeds the stress threshold after a specific time, LIZA offers assistance through stress coping strategies. Among the twelve coping approaches classified by Skinner and colleagues (Skinner et al., 2005), the strategies “Isolation”, “Delegation”, “Opposition”, “Negotiation”, “Escape”, “Support seeking” seem to be not goal-oriented in the context of learning. The approaches “Information seeking”, “Self-reliance”, and “Accommodation” could be deployed in an (online) learning environment. “Self-reliance” methods could be mindfulness or autogenic training. These strategies require a longer period. Adopting those "Self-reliance" strategies, the pedagogical agent may prepare appropriate video sequences for mindfulness or autogenic training. "Accommodation" strategies could be telling jokes or showing a video. These strategies are adopted in the current version. The first one distracts the user by telling jokes, the second one shows a video with relaxing content. The user decides whether it is necessary to start the offered coping process and how long the strategies are used. If the stress level is significantly reduced below the threshold, pedagogical agent LIZA proposes the continuation of the learning process (see Figure 4).

2.5 A Dialogue Design

As mentioned before, it was necessary to alter the pedagogical intervention process of the original pedagogical agent LIZA for analyzing the RMSSD data accordingly (see Figure 5). First, an inquiry for the declaration of consent for monitoring the heart rate was added to the greeting phase. After that, the user is requested to apply and activate the wristband and start the mobile transmission application. If a specific time interval is requested by the pedagogical agent, the suitable NN heartbeat intervals will be selected based on the time stamp and the RMSSD of these values will be determined. After the learning phase, the performance of the learner is evaluated and the final score is reported to the user (Figure 5).

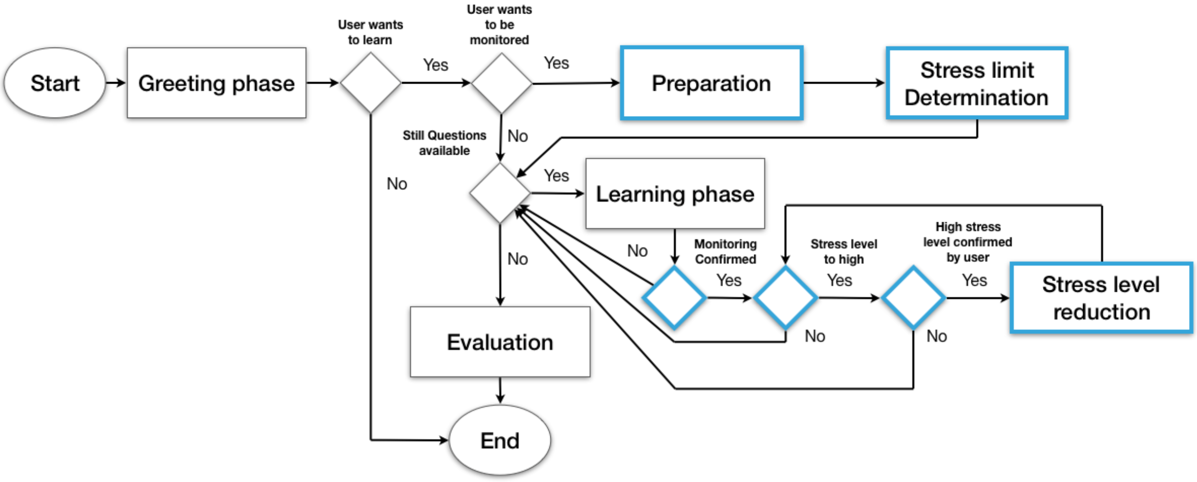


Figure 5. A Dialogue-based Intervention Process.

3. Evaluation

The goal of the evaluation study is to determine the effectiveness and benefits of the pedagogical agent LIZA that was extended with the capability of measuring HRV and detecting the critical stress level of learners. Amongst others, the following hypotheses are examined:

1. The RMSSD is a suitable indicator for determining the learner's critical stress threshold.
2. The proposed stress reduction strategies lead to the relaxation of learners.

3.1 Design

For the evaluation study, 34 participants (10 males, 24 females) aged between 21 and 59 (mean 31 ± 11 years) were acquired and assigned to a test or a control condition by random. The control condition was required to evaluate the effect of the proposed stress reduction strategies (which is Hypothesis 2). The study was conducted in a quiet environment under the supervision of the project leader. In both conditions, each participant was asked to use the pedagogical agent to perform two stress tests, each with a different level of difficulty, to determine the learner's stress threshold (cf. Section 2.3).

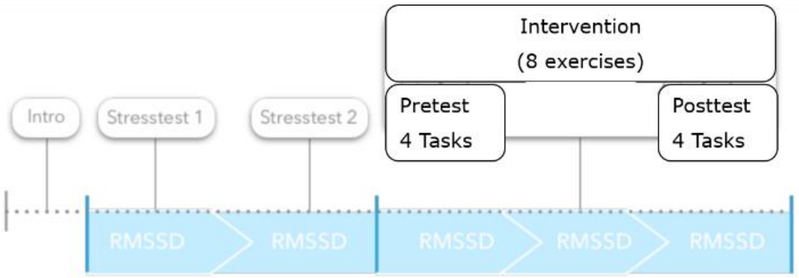


*Figure **6***. Experiment Procedure.

Figure 6 summarizes the experiment procedure. In the first phase, each participant is invited to take a seat in front of a notebook. He/she is informed and explained about the procedure of the experiment by a study supervisor. After signing a consent form and answering the demographic form (gender, age, profession, medical-relevant aspects including heart disease, depression, alcohol addiction), an Empatica E4 wristband is put on the wrist of the participant. Data transfer between the wristband and the server is activated and calibrated.

Next, in the second phase, each participant is asked to use the developed pedagogical agent to carry out the two stress tests. The individual stress threshold is determined in this phase.

In the third phase, in which learning should take place, participants can start to solve reasoning tasks provided by the pedagogical agent. The tasks are grouped into two blocks: a pretest and a posttest, each is composed of four tasks. The experiment is designed with a pretest and a posttest to determine any learning effect. After the first block, RMSSD is calculated. Only the test condition has a stress-coping phase between the two task blocks if the current RMSSD exceeds the threshold. After the stress-coping phase, participants are asked to solve again 4 tasks. The reasoning task categories of the pretest and posttest are the same, but the tasks are different. In the end, each participant gets an evaluation of how successful the tasks have been solved. For every phase of the experiment procedure (pretest, stress-coping phase, posttest), RMSSD is calculated so that the development of the indicator can be followed. In addition, the participants are required to self-report their current state of mental load after each measurement cycle using a short questionnaire KAB (Berth, 2003). The self-report is required to carry out a concurrent validity for proving that RMSSD is a relevant instrument for measuring stress (Adams et al., 2014). There exist different stress self-report instruments for different purposes, e.g., Aerospace (Draycott & Kline, 1996). To measure stress in the context of learning, a suitable self-report is required. The KAB questionnaire has six adjective-pairs representing six questions: 1) tense-calm, 2) anxious-unconcerned, 3) worried-carefree, 4) restless-relaxed, 5) skeptical-trusting, 6) uncomfortable-comfortable. The rating for each question varies on a scale between 1 (the left adjective) and 6 (the right adjective). The average KAB index is calculated by the sum of ratings for all questions divided by six. If $KAB_{index} \leq 3$, that means, the subjective perception of cognitive load is in the stressful state, otherwise if $KAB_{index} > 3$, the subject is in the non-stressful state, i.e., comfortable state. The experiment procedure of the control condition does not include the stress-coping phase.

3.2 Results

Hypothesis 1

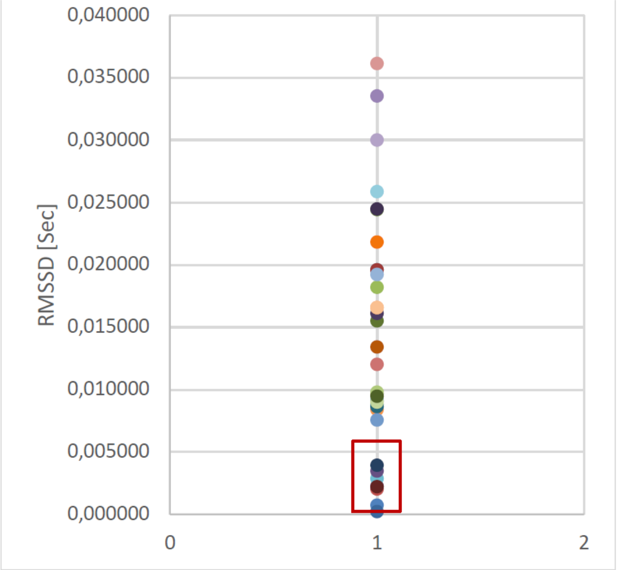


Figure 7. RMSSD Differences between Two Stress Tests.

Figure 7 shows the difference between the RMSSD values of the two stress tests. On the Y-axis, the absolute difference between the RMSSD values of each participant is displayed. The design of the stress tests assumes that through the difficulty of the test, each participant may have a range between normal cognitive load and stressful cognitive load and this range can be determined using RMSSD as a stress indicator. Statistical results show that the absolute difference between RMSSD values of two stress tests is 0.0136 seconds (=13.6 milliseconds) on average over all participants and the difference is significant (p= 0.0079) at a significance level of 0.05. That is, RMSSD can be used to determine individual stress thresholds in real-time, although the RMSSD difference between the stress tests is small. Such a small RMSSD difference could be explained by two possible reasons. The first case could be that both tests induce a small cognitive load. As a result, the stress-coping strategy was introduced in a state with a small cognitive load. The second case could be that both stress tests induce a high cognitive load. This case, ideally, would not influence the stress-coping phase. However, for individual learners, a high cognitive load could last a long time.

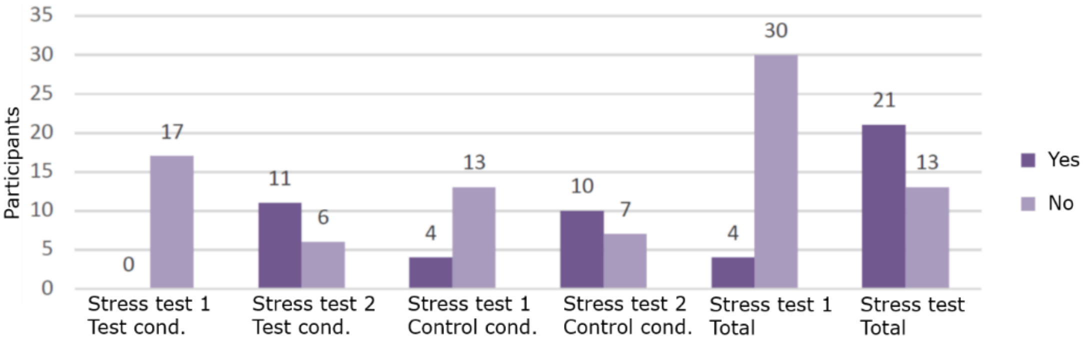


Figure 8. Participants who rated the Tests with $\text{KAB}_{\text{index}} \leq 3$ (i.e. stressful).

Self-reports can be used to concurrently validate the results of RMSSD. For this purpose, the KAB index has been considered as an instrument to measure subjective stress state. Only four of thirty-four participants self-report a KAB index less than 3 after the first stress test, i.e. they felt stressed. After the second stress test, twenty-one of thirty-four participants rated their KAB index less than 3, i.e., the second stress test induces stress (see Figure 8). The KAB index difference between the first and the second stress tests is significant (p=4.08199 E-10). This result corresponds to the significant RDSSM difference between the two stress tests. Therefore, Hypothesis 1 can be confirmed.

Hypothesis 2

Table 1. *Results after Pretest, the 1st Stress Coping and the 2nd Stress Coping Phases*

	After Pretest				**After 1st Stress Coping**					**After 2nd Stress Coping**				
Nr.	RDSSM threshold exceeded	Stress confirmed	KAB≤3	KAB≤4	Tendency RDSSM	Tendency KAB	KAB>3	KAB>4	RDSSM threshold exceeded	Stress confirmed	Tendency RDSSM	Tendency KAB	KAB>3	RDSSM threshold exceeded
01	Y	Y	Y	Y	↓	↑	Y	Y	Y	Y	↓		Y	Y
02	Y	Y	Y	Y	↑	↑	Y	Y						
04	Y	Y	Y	Y	↓	↑			Y	Y	↓	↑	Y	Y
07	Y	Y	Y	Y	↑	↑	Y							
08	Y	Y		Y	↑	↑	Y	Y						
09	Y	Y		Y	↑	↑	Y	Y						
12	Y													
13	Y	Y		Y	↑	↑	Y	Y	Y					
14	Y	Y			↑	↑	Y	Y	Y					
16	Y	Y	Y	Y	↓	↑	Y		Y					
17	Y	Y		Y	↓	↑	Y		Y	Y	↑	↑	Y	Y

To test Hypothesis 2, RDSSM values and KAB indexes before and after the stress-coping phase are compared. Table 1 summarizes results after the pretest, after the first and the second stress-coping phases. In each phase (represented by each bound column), the participants whose RDSSM exceeded the RDSSM threshold are listed and they had the opportunity to confirm their stress state subjectively upon a question asked by the pedagogical agent. After solving four reasoning tasks of the pretest, there were eleven participants in the test condition whose RDSSM exceeds the stress threshold after solving the four tasks of the pretest, and ten of them confirmed their stress state subjectively. Thus, the determination of the stress threshold based on RDSSM was reasonable. Among those ten participants, five of them had a KAB index less equal to 3. If the KAB threshold for stress were defined at scale 4, there would be nine participants exceeding the threshold. In the stress-coping phase, each participant whose RDSSM exceeded the threshold has two chances to apply a stress-coping strategy. After applying the first stress-coping strategy, ten participants (who have confirmed their stress state subjectively) rated higher KAB index than after the pretest. That means these ten participants could reduce their stress. The difference of KAB indexes between the first stress-coping strategy and the pretest is significant with p=0.0034. Among these ten participants, nine of them had a KAB index greater than 3, i.e. they achieved a relaxed state (see Table 1, 2nd bound column). Considering the change of RMSSD of these ten participants, we can learn from statistics that only six participants had an increase of RMSSD which is an indicator of stress recovery. Four of these six participants had RMSSD values below their stress threshold and thus could finish their stress-coping phase. Analyzing individual participants, we can learn that participant #16 has opposite tendencies of RMSSD (↓) and KAB (↑). It is noted that the RMSSD value of participant #16 lies on the threshold, thus, the opposed tendencies of RMSSD and KAB can be accepted. Also, participant #17 showed at the end of the stress-coping phase increasing KAB and RMSSD. Thus, this could be assumed that a longer application of stress-coping strategy results in better stress recovery. Especially, this tendency is visible after applying the second stress-coping strategy. It is also worth investigating individual participants (#01, #04, #17) whose RMSSD still exceeds the stress threshold even after the second stress-coping strategy was applied. The reason is that the range of RMSSD between the two stress tests of these participants was too small, and thus the calculated stress threshold could be very sensitive. Also, participant #13 had this problem. The RMSSD difference between the two stress tests of this participant was 7.5 milliseconds which is the smallest range among other participants.

The second hypothesis, whether the stress reduction phase leads to the relaxation of the learner, could be partially confirmed. 90% of the participants stated in a self-assessment a recovery, which indicates the effect of the applied stress reduction strategies. But only in nearly 50% of the cases, the RMSSD also falls below the threshold. Possible reasons for that could be deficits in stress threshold determination, insufficient choice of strategies, or insufficient application time. Taking t-test analysis of RMSSD, the RMSSD difference between the stress-coping phase and the pretest is 5.9 milliseconds and is not statistically significant (p= 0.2903).

4. Conclusions and Future Work

This paper has demonstrated the integration of physiological factors in technology-enhanced learning environments using wearable sensors. The contribution of the study presented in this paper is two-fold. It helps researchers to choose RMSSD as a relevant metric for determining the individual stress threshold level of learners in computer-supported learning environments. Second, it suggests embedding feasible relaxation strategies into learning systems. Due to the limit of this paper, data regarding pretest and posttest have not been analyzed. In near future, the impact of stress-coping strategies in the context of learning needs to be analyzed.

References

Adams, H., Cervantes, P., Jang, J., Dixon, D. (2014). Evidence-Based Treatment for Children with Autism. In Practical Resources for the Mental Health Professional, Chapter 25 - Standardized Assessment, D. Granpeesheh, J. Tarbox, A. C. Najdowski, J. Kornack, (eds.), Academic Press, 501-516. DOI:10.1016/B978-0-12-411603-0.00025-2.

Berth H. 2003. KAB. In: Berth, H. & Balck, F. (eds.). Psychological Tests for medical professionals (German: Psychologische Tests für Mediziner). Berlin: Springer Verlag, 148-149.

Castaldo, R., Melillo, P., Bracale, U., Caserta, M., Triassi, M. & Pecchia, L. (2015). Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis. Biomedical Signal Processing and Control, 18, p. 374.

Chatti, M., Lukarov, V., Thüs, H., Muslim, A., Yousef, A. M. F., Wahid, U., Greven, C., Chakrabarti, A., Schroeder, U. (2014). Learning Analytics: Challenges and Future Research Directions. Eleed.

D'Mello, S. A. K. (2017). The Affective Computing Approach to Affect Measurement. Emotion Review. DOI:10.1177/1754073917696583

Draycott, S. G. & Kline, P. (1996). Validation of the AGARD STRES Battery of Performance Tests. Human Factors. 38(2), 347-361. DOI:10.1177/001872089606380214

Empatica Inc. (2016). Utilizing the PPG/BVP Signal. https//support.empatica.com/hc/en-us/articles/204954639-Utilizing-the-PPG-BVP-signal

Endukuru, C.K. & Tripathi, S. (2016). Evaluation of cardiac responses to stress in healthy individuals - a noninvasive evaluation by heart rate variability and stroop test. International Journal of Scientific Research, 5(7).

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. International Journal of Technology Enhanced Learning, 4, p. 304-317. DOI: 10.1504/IJTEL.2012.051816.

Giannakos, M. N., Sharma, K., Papavlasopoulou, S., Pappas, I. O., Kostakos, V. (2020). Fitbit for learning: Towards capturing the learning experience using wearable sensing, International Journal of Human-Computer Studies, 136, 102384, ISSN 1071-5819. DOI:10.1016/j.ijhcs.2019.102384.

Gjoreski, M. & Gjoreski, H. (2017). Monitoring Stress with a Wrist Device Using Context. Journal of Biomedical Informatics, 73, 159-170.

Kim, H. G., Cheon, E. J., Bai, D. S., Lee, Y. H. & Koo, B. H. (2018). Stress and Heart Rate Variability: A Meta-Analysis and Review of the Literature. Psychiatry Investigation, 15(3), p. 235.

Kotter, T., Wagner, J., Bruheim, L., & Voltmer, E. (2017). Perceived Medical School stress of undergraduate medical students predicts academic performance: An observational study. BMC Medical Education, 171, p. 256. https://www.ncbi.nlm.nih.gov/pubmed/29246231

Le, N.T. & Wartschinski, L. (2018). A Cognitive Assistant for improving human reasoning skills. International Journal of Human-Computer Studies, 117, 45-54, ISSN 1071-5819. DOI: 10.1016/j.ijhcs.2018.02.005.

Li, Q., Xue, Y., Zhao, L., Jia, J., Feng, L. (2017). Analyzing and identifying teens stressful periods and stressor events from a Microblog. IEEE Journal of Biomedical and Health Informatics 21, 1434 – 1448. DOI:10.1109/JBHI.2016.2586519

Li, Z., Snieder, H., Su, S., Ding, X., Thayer, J.F., Treiber, F.A. & Wang, X. (2009). A longitual study in youth of heart rate variability at rest and in response to stress. International Journal Psychophysiology, 73(3), 212-217.

Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. EDUCAUSE Review, 46(5), 31–40.

Lyu, Y., Luo, X., Zhou, J., Yu, C., Miao, C., Wang, T., Shi, Y. & Kameyama, K. (2015). Measuring Photoplethymogram-Based Stress-Induces Vascular Response Index to Assess Cognitive Load and Stress. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 857-866.

Mourenas, D., Zorila, M. & Meinders, E. (2018). Analysis of physiological signals for recognition of stress. https://mentechinnovation.eu/wpcontent/uploads/2018/10/Analysis_of_physiological_signals_for_recognition_of_stress.pdf

Pascoe, M. C., Hetrick, S. E. & Parker, A. G. (2020) The impact of stress on students in secondary school and higher education, International Journal of Adolescence and Youth, 25:1, 104-112, DOI:10.1080/02673843.2019.1596823

Punita, P., Saranya, K., Chandrasekar, M. & Kumar, S. (2016). Gender difference in heart rate variability in medical students and association with the level of stress. National Journal of Physiology, Pharmacy and Pharmacology, 6(5).

Rüdian, L., Gundlach, J., Vladova, G., Pinkwart, N. & Kazimzade, G. (2019). Predicting culture and personality in online courses. In Proceedings of the Workshop SLLL@AIED 2019.

Sammito, S., Thielmann, B., et al. (2014). S2k-Leitlinie: Nutzung der Herzfrequenz und der Herzfrequenzvariabilität in der Arbeitsmedizin und Arbeitswissenschaft. AWMF online.

Schneider, M., Jacobs, D. W., et al. (2003). Cardiovascular Hemodynamic Response to Repeated Sental Stress in Normotensive Subjects at Genetic Risk of Dypertension: Evidence of Enhanced Reactivity, Blunted Adaption and Delayed Recovery. Journal of Human Hypertension, 17(12), 829-840.

Selye. H. (1976), Stress in health and disease, Boston: Butterworth (Publishers) Inc.

Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Shum, S. B., Ferguson, R., et al. (2011). Open Learning Analytics: an integrated & modularized platform. Proposal to design, implement and evaluate an open platform to integrate heterogeneous learning analytics techniques.

Skinner, E. A., Edge, K., Altman, J. & Sherwood, H. (2003). Searching for the Structure of Coping: A Review and Critique of Category Systems for Classifying Ways of Coping. Psychological Bulletin, 129(2), p.245.

Taelman, J., Vanduput, S., Vleminx, E., Spaepen, A. & van Huffel, S. (2011). Instantenous changes in heart rate regulation due to mental load in simulated office work. European Journal of applied Physiology, 1111, 1497-1505.

Tharion, E., Parhasarathy, S., & Neelakantan, N. (209). Short-term heart rate variability measures in students during examinations. The National Medical Journal of India, 22(2), 63-66.

Visnovcova, Z., Mestanik, M., Javorka, M., Mokra, D., Gala, M., Jurko, A., Calkovska, A. & Tonhjzerova, I. (2014). Complexity and time asymmetry of heart rate variability are altered in acute mental stress. Physiological Measurement, 35, 1319-1334.

Zangroniz, R., Martinez-Rodrigo, A. et al. (2018). Estimation of Mental Distress from Photoplethysmography. Applied Science, 8(69), 1-15.

A Machine Learning Approach to Estimating Student Mastery by Predicting Feedback Request and Solving Time in Online Learning System

Kannan N[a*], Charles Y. C. YEH[a], Chih-Yueh CHOU[b] & Tak-Wai CHAN[a]
[a]*Graduate Institute Of Network Learning Technology, National Central University, Taiwan*
[b]*Department of Computer Science and Engineering, Yuan Ze University, Taoyuan, Taiwan*
*kannannataraj@g.ncu.edu.tw

Abstract: One of the most significant challenges for computers in education is the capacity to provide intelligent and adaptable learning systems to meet the real needs of students. In order to create efficient adaptive or personalized mechanisms for educational content, student models are proposed to estimate the actual knowledge or mastery level of students. Some earlier student models were proposed to estimate student mastery based on the correctness (e.g., correct or incorrect) of responses, feedback request, and solving time using classical Markov process and logistic regression models. In particular, these models were applied to predicting student future correctness, feedback request, and solving time (i.e., on the next question).The advent of increasingly large-scale datasets has turned Machine Learning (ML) methods such as conventional machine-learning algorithms and deep learning models for prediction into competitive alternatives to classical Markov process and logistic regression models. In addition, prediction by ML methods has numerous advantages such as interpretability, good accuracy, ease of maintenance, less execution time, and appropriately handling of missing data. Moreover, recent studies exhibit the significant achievement of ML prediction methods for estimating students' performance and mastery using learning log data (i.e., correctness, feedback request level, solving time, etc.). Hence, it is reasonable to use ML methods to estimate student mastery by predicting the feedback request level and solving time. This study analyzed the data logged by an online learning system called Math-Island, which teaches elementary level mathematics by incorporating game mechanisms and scaffolding feedback. Machine-learning regression methods such as Multiple Linear Regression (MLR), Support Vector Regression (SVR), Random Forest Regression (RFR), Extra Trees (ET), and Gradient Boosting Regression (GBR) were applied. The results showed that RFR and GBR were found to outperform other models to predict future feedback request level and solving time. The results lead to several future works. First, incorporating ML predictive models into Math-Island tutoring system to identify the individual student's actual needs and; reduce learning loss substantially. Second, it drives to effectively build a more efficient adaptive mechanism within the current session to utilize students' active learning time.

Keywords: Machine learning, prediction, regression, online learning system.

1. Introduction

Estimating students' learning performance and content knowledge or mastery is a long-standing practice in the education field, which contributes to supporting and enhancing learning achievements. Mastery-oriented goals focus students' attention on achievement based on intrapersonal learning standards; performance goals focus on achievement based on normative or comparative performance standards. Ansems et al. (2019) describes that mastery goals focus on developing competence and mastering a task. In contrast, performance goals focus on the demonstration of competence and outperforming others. Research evidence suggests a mastery goal orientation that promotes a motivational pattern likely to promote long-term and high-quality involvement in learning (Tuominen, Juntunen, & Niemivirta, 2020). Moreover, studies show that mastery-oriented goals consistently lead to intrinsically motivated, self-regulated learning and promote comprehension (Ansems et al., 2019; Caniëls, Chiocchio, &vanLoon, 2019). On the other hand, researchers suggested one-to-one technology (Chan et al., 2006) through which every student is equipped with a device to learn in school or at home seamlessly. Online learning systems (Jin, 2020; Yeh, Cheng, Chen, Liao, & Chan, 2019) successfully encompass this feature: they teach skills, such as algebra, numeric operation, geometry, computer programming, or medical diagnosis, using mastery-oriented goal principles and provide learners with individualized feedback and materials adapted to their level of understanding (Romero, Hernández, Juola, Casadevante, & Santacreu, 2020). Studies have demonstrated that online learning has gained much attention in recent years. However, it needs improvements and incorporates new technologies as per large-scale datasets (Ogdanova, Tova, & Vetaeva, 2014) to make an efficient adaptive and flexible learning context.

Mastery estimation and prediction is used to identify what a student will do or know at the end of an instructional unit (Bälter, Zimmaro, & Thille, 2018). Mastery estimation has been varying in line with the evolution of tutoring and learning methods. The well-known student models, Bayesian knowledge tracing (BKT) by Corbett and Anderson (1994), and performance factor analysis (PFA) by (Pavlik, Cen, & Koedinger, 2009), have been widely used to estimate current student knowledge mastery. BKT was proposed during the 1990s; BKT predicts student mastery via probabilities that include four parameters per knowledge component. PFA uses logistic regression to predict mastery as the output of the learned or unlearned state. These models predict student mastery based on the correctness of responses. In particular, these models attempt to predict students' future correctness (i.e., on the next question). However, there are few significant limitations to these methods; for example, BKT assumes no downsides and cannot forget once students learn the skill. In contrast, Bälter (2018) mentioned that once learned, there are also downsides with continued repetition and might even be detrimental under certain conditions. Moreover, previously reported results to show that BKT cannot handle missing data patterns for prediction (Gervet, Koedinger, Schneider, & Mitchell, 2020). Furthermore, Pavlik et al. (2009) implemented the PFA algorithm in Excel, and using Excel is not feasible for large-scale datasets and on-time supports. Moreover, both models were not fully able to account for more rapid shifts in student performance, especially in cases where a student struggles early but goes on to drastically improve their performance (Slater & Baker, 2019). Therefore, these limitations have turned machine-learning methods for prediction into competitive alternatives to classical Markov process and logistic regression models. Furthermore, Machine-learning methods such as Random Forest, linear regression, feed-forward neural network, and deep learning models such as Deep Knowledge Tracing were all shown to deliver superior results to BKT and PFA on their own (Mao, Lin, & Chi, 2018; Piech et al., 2015).

Machine-Learning (ML) methods can be applied throughout science, technology, and commerce, leading to more evidence-based decision-making across various fields, including education, health care, manufacturing, financial modeling, policing, and marketing. It mainly has a considerable impact on the educational field, especially for estimating, tracing and predicting students’ mastery and performance (Imran, Latif, Mehmood, & Shah, 2019; Sokkhey & Okazaki, 2020). In the technology-pervasive world, as a student is working toward a solution, the system keeps track of his or her actions and provides feedback to help the student progress (Myneni, Narayanan, Rebello, Rouinfar, & Pumtambekar, 2013). A study (Lai & Lin, 2015) found that students received immediate, elaborative, text-based feedback led to more effective learning and higher motivation. Results suggest that immediate feedback from computer-based learning tasks benefit both high and low prior knowledge students, with low prior knowledge students exhibiting more significant gains (Razzaq, Ostrow, & Heffernan, 2020). Solving time is based on correctness, which is collected separately within each attempt. Student solving time has been mainly used to assess student learning because it can indicate how active and accessible student knowledge is (Mao et al., 2018). For example, it has been shown that solving time reveals student mastery, and has been suggested as an indicator of student engagement in answering questions as well as an important factor for predicting motivation in an e-learning environment (Schnipke, 2002).

Therefore, in this work, we used both feedback request level and solving time as feature variables for prediction using ML methods to estimate students' hidden mastery level. ML methods thrive on large datasets, create assumptions about the data, and allow them to use non-normally distributed input variables. Moreover, the ML approach has several advantages: interpretability, accuracy, ease of maintenance, adequate execution time, and appropriately handling missing data, unlike BKT, PFA. In addition, we evaluate models only in terms of the accuracy of their predictions—the resulting best models to employ for carrying out adaptive pedagogical strategies and further advanced personalized learning for future works.

2. Machine Learning Methods in Student Performance Prediction

Students' learning progress often refers to student knowledge as a latent variable (Corbett &Anderson, 1994). Estimating actual knowledge of the content and providing enough practice opportunities before moving on to the next level are challenging for educators, particularly in an online learning environment or distance learning and flexible learning context. Pelánek (2015) mentioned that the over-practice (the practice of items that the student already mastered, i.e., “wasted time” of students) and under-practice (a missing practice that is necessary for mastery of a topic and further progress) are the broader problems for learners, which is driven them to loss of learning. In a study of high school students participating in the Optimized Cognitive Tutor geometry curriculum, it was found that 58% out of 4102 practices and 31% of 636 exercise questions were done after the students had reached mastery (Cen, Koedinger, & Junker, 2007). However, it is possible to reduce study time without the loss of learning (Bälter et al., 2018). A prior study conducted with 265 individuals shows that over half the individuals are expected to need less than five practice opportunities to reach mastery. In addition, over 40 students require at least 15 practice opportunities, and over 30 of those require 20 or more to reach mastery (Lee & Brunskill, 2012). Moreover, Beck and Gong (2013) found that 69% of students in Cognitive Algebra Tutor (CAT) and 62% of students in ASSISTments have mastered the skill after ten practice opportunities.

The findings above suggest that the optimal number of questions needed to prevent students from running out of insufficient practice opportunities before they have mastered a skill, simultaneously not spending more time and resources on an already mastered skill (Bälter et al., 2018). Slater &Baker (2019) recommend that mastery or knowledge estimation is a valuable technique for quickly identifying student's needs likely to wheel-spin (Beck & Gong, 2013) and providing additional scaffolding or support for their learning. Therefore, estimating student mastery of skill after solving the optimal number of questions could alleviate the loss of learning for high-low achieving students.

Machine learning is widely used for prediction problems, especially in the education field. In other words, machine-learning methods showed good performance than other statistical methods. For instance, recent studies have demonstrated good accuracy in predicting student performance in contexts of solving problems in Intelligent Tutoring Systems (ITSs) or completing courses in a classroom or in Massive Open Online Courses (MOOC) platforms (Jin, 2020; Mao et al., 2018). In most of the previous literature (Table 1), many researchers have approached to find the best algorithms for estimating future performance based on predicting correctness, final grade, dropout status, and final exam scores; moreover, most studies are proposed classification—few used regression methods due to the nature and relationship of each input variable with the output variable. In addition, the studies mentioned earlier have used longer timescale outcome prediction due to the dataset and variables (i.e., semester result, score, and grade). Different models have different sensitivities to the type of predictors in the model; how the predictors enter the model is also important.

Table 1. *Summary of Recently Proposed ML Methods in ITSs for Predict Student Performance*

Algorithm	Reference	Study focus	Input Variables	Output Variable
Naïve Bayes, Generalized Linear Model, Logistic Regression, Deep Learning, Decision Tree, Random Forest, and Gradient Boosted Trees.	Abidi, Hussain, Xu, &Zhang, 2018	Academic performance of students	Attempt_count, hint_total, overlap_time, response type (correct or incorrect)	Final grade
Artificial Neural Networks, Support Vector Machines, Logistic Regression, Naïve	Hussain, Zhu, Zhang,	Student’s difficulty during the next session	Average time, total number of activities, average idle time, average number of	Grades

bayes Classifiers and Decision trees	Abidi, &Ali, 2019		keystrokes and total related activity	
Logistic Regression	Asselman, Khaldi, &Aammou, 2020	Prior required scaffolding items to predict future student performance	Prior scaffolding	Correctness
Multilayer Perceptrons , Sequential Minimal Optimization of Support Vector Machine, Logistic Regression, and Random Forest.	Sokkhey &Okazaki, 2019	Student performance in mathematics	Domestic, academic, and attitudinal information describing each student,	Grade
Linear Regression	Rong &Bao-Wen, 2018	Students at risk of failure	Prior term's student clicker data and final exam scores, high school Grade Point Average (GPA), high school GPA, gender, age, socioeconomic status	Final exam score
Support Vector Machine, Random Forest, and Extra Trees	Jin, 2020	Students' early dropout status in MOOC.	Clickstream data	Drop out status

Accordingly, this study applies multiple regression algorithms such as Multiple Linear Regression (MLR), Support Vector Regression (SVR), Random Forest Regression (RFR), Extra Trees (ET), and Gradient Boosting Regression (GBR) to predict numerical output variables (i.e., feedback request level and solving time) to estimating student mastery, which happens in a specific time (i.e., shorter timescale). Moreover, in this paper, we evaluate models only in terms of the accuracy of their predictions.

3. Method

This study's data were collected from students in Taiwan who used the "Math-Island online learning game" during 2016-2019. Math-Island is an online learning system that incorporates gamified knowledge map of the elementary mathematics curriculum (Yeh et al., 2019). The Math-Island game targets the mathematics curriculum of elementary schools in Taiwan, mainly containing the four domains: numerical operation, quantity and measure, geometry, and statistics and probability (Figure 1 a). Each domain contains gamified knowledge map of concept units for students to learn mathematics by units (Figure 1 b). Every unit has a different quantity of tasks (Figure 1 c), and each task has several problem-solving questions (Figure 1 d). Students can freely choose the learning path according to their interests, and if students face difficulty while learning the skill, the system will provide scaffolding feedback.

The system collects many data when students answer problem-solving questions, including problem-related information, and instructional assistance, e.g., student id, number and level of feedback request, solving time, and total log time on individual questions. We use Pearson's correlation coefficient (Figure 3) method to choose feature variables; the plot showing the positive correlation coefficient between the input variables and output variable, which determined that the regression algorithms are more appropriate for prediction. Therefore, this study applied machine-learning regression methods to predict feedback request level and solving time using the following variables:

Feedback-level is the total number of feedback or attempts students used each question (minimum 0 - maximum 3). (0) Feedback (such as "correct"), (1) Feedback (such as "incorrect"), (2) Feedback + scaffolded solution (how to answer), (3) All three services with the answer (Figure 2).

Solving time is how long it takes for students to finish each question, excluding every feedbacks time if students answered the questions incorrectly. We used these variables to build two different prediction models accordingly. Variables collected from the first four questions are applied to predict the variable of the fifth question (Table 2).

Table 2. *Description of Input Variables and Output Variable*

Independent variables				Dependent variable
Question 1 x_1	Question 2 x_2	Question 3 x_3	Question 4 x_4	Question 5 Y
Feedback request level	Feedback request level	Feedback request level	Feedback request level	Feedback request level
Solving time	Solving time	Solving time	Solving time	Solving time

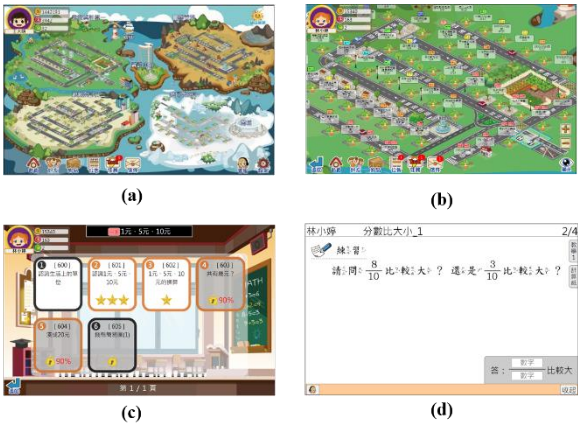
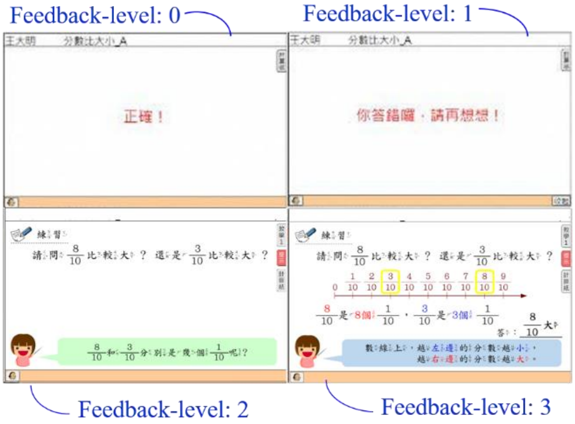


Figure 1. Math-Island online learning system and task screen



正確！

你答錯囉，請再想想！

Figure 2. Feedback-levels

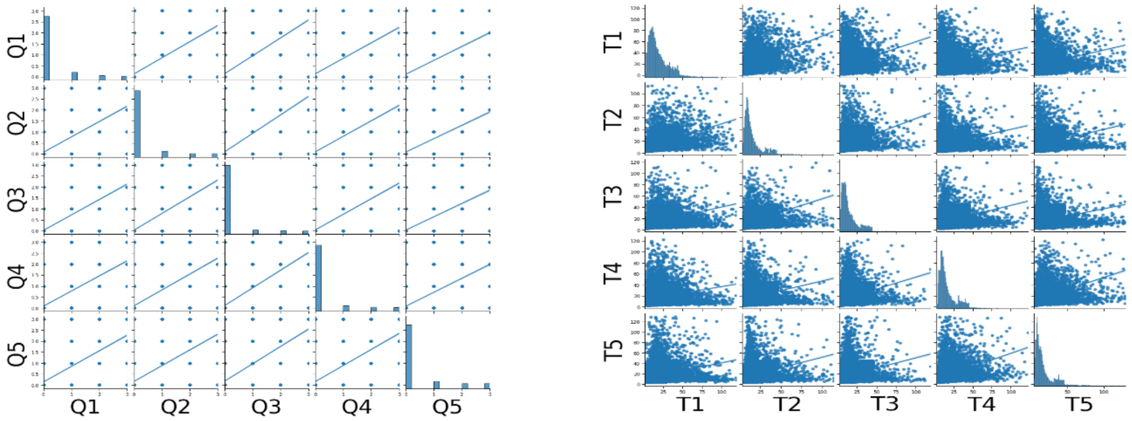


Figure 1. Relationship between Input variables and Output Variables. (Pair plot distribution shows the positive correlation coefficient between the input variables and output variable)
Keys: Q1-Q5 – feedback levels of Question 1 to Question 5; T1-T5 – Solving time of Question 1 – Question 5

Our approach is to estimate students’ mastery by predicting their feedback request level and solving time. We assume that the predictive output of feedback request level (0-1) and solving time (5-15) is the threshold of mastery (Figure 4); nevertheless, more exploratory study is needed to establish our findings here. However, this study aims to estimate the regression model with coefficients of c, w0 + w1 x1 + w2 x2 + … + wn and fit the training data with minimal squared error and predict the output y for each to find the best prediction model.

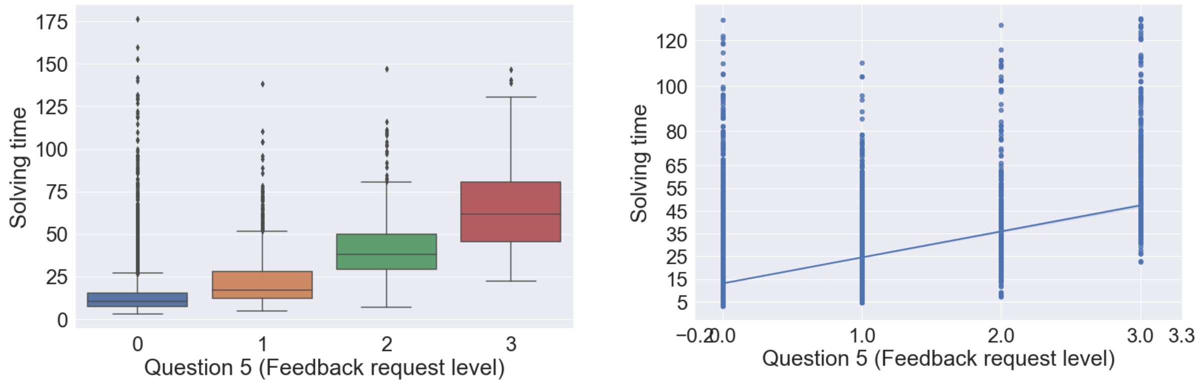


Figure 2. Distribution of utilized feedback request level according to the solving time on question five (left: the box plot distribution of feedback request level according to solving time, right: the scatterplot of two variables with 95% confidence interval of regression line)

4. Data Collection and Analysis

For the initial stage of this study, we used data collected from one task unit in the numerical operation domain in Math-Island, where there are 50 task units, and the level of difficulty is identical for all questions. The data contains 13168 answer records, including student id, mission id, feedback request level, and solving time of each question. After removing outliers (e.g., less than 1.5 sec, more than 140 sec, and null values), 12034 records were chosen for this study.

Table 3. *Dataset Description.*

Sl.no	Q1	Q2	Q3	Q4	Q5	T1	T2	T3	T4	T5
0	3	1	0	0	3	62.811	35.696	10.104	10.527	79.831
1	3	0	0	0	0	33.524	9.343	11.446	11.017	11.524
2	3	0	0	0	1	69.260	13.187	25.633	18.066	33.733
3	0	0	0	0	0	9.065	10.573	26.696	8.785	5.783
…	...	...	...	...	...	...	...	...	...	...
12033	2	2	3	2	2	35.000	39.000	35.000	35.000	45.000
min	0	0	0	0	0	1.68	2.80	2.56	2.61	3.17
max	3	3	3	3	3	119.23	113.91	119.38	122.46	129.66
count	12034	12034	12034	12034	12034	12034	12034.	12034	12034	12034
mean	0.363	0.322	0.276	0.329	0.394	20.27	16.103	16.031	17.817	17.562
SD	0.79	0.78	0.74	0.79	0.84	15.14	12.67	12.24	13.94	15.08

5. Model Evaluation

The dataset has split into a training dataset and test dataset within 75-25 ratio. We use two standard regression metrics Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), to evaluate applied methods' performance. The evaluation included a comparison between the prediction results of the models (Table 4). Note that both of these metrics measure error, so a smaller number indicates a better predictive model. [1]

The simple baseline was measured by central tendency measurement that used the global mean of Y and then calculated MAE and RMSE of the mean by reducing the y_test data. The baseline was measured to infer the performance of each model.

[1] This study used Python 3.6 to analyze the data with the following kits: the Numpy suite for data collation, Pandas, Matplotlib, the SciPy kit for data visualization, Seaborn, and the Scikit-learn kit.

Table 4. *Model Evaluation and Comparison between the Models*

Algorithms	Feedback request level		Solving time	
	MAE	RMSE	MAE	RMSE
Baseline	**0.606**	**0.827**	**10.418**	**14.906**
Multiple Linear Regression	0.335	0.543	7.628	12.856
Support Vector Regression	0.384	0.550	8.065	14.763
Random Forest Regression	0.333	0.530	7.361	12.756
Extra Trees Regression	0.335	0.530	7.461	12.894
Gradient Boosting Regression	0.352	0.529	7.030	12.342

Table 5. *Comparison of Actual (y_test) and Prediction (y_pred) Values*

	Feedback request level (y_pred)						Solving time (y_pred)					
	y_test	MLR	SVR	RFR	ET	GBR	y_test	MLR	SVR	RFR	ET	GBR
count	3009	3009	3009	3009	3009	3009	3009	3009	3009	3009	3009	3009
mean	**0.37**	0.39	0.30	**0.38**	**0.38**	**0.38**	**17.14**	**17.37**	12.96	17.86	18.00	**17.32**
SD	0.82	0.62	0.66	0.64	0.64	0.63	14.90	7.86	3.41	9.67	9.63	8.61
min	0	0.10	**-0.08**	0.00	0.00	0.10	3.33	7.01	3.67	4.26	4.37	4.29
max	3	2.84	2.90	2.78	3.00	2.52	126.83	59.84	23.37	91.06	83.54	63.91

Table 4 shows the performance comparison among each model to predict the final question's feedback request level and solving time. Table 5 shows the comparison of each model's prediction outputs with the actual value. The result of the five models has outperformed baseline MAE and RMSE. However, with comparing each model for feedback request level, RFR (0.333), MLR (0.335), and ET (0.335) seem to have minimal MAE than other models; on the other hand, GBR has minimal RMSE (0.529). In addition, for solving time, GBR (MAE: 7.030, RMSE: 12.342) outperformed other models.

MLR (MAE: 0.335, RMSE: 0.543) for feedback request level prediction performed better than baseline, and it produced the closest prediction output with the actual value; standard deviation (SD) also is significantly less than other predictors. On the other hand, for solving time, MLR (MAE: 7.628, RMSE: 12.856) performed better than RFR, ET, SVR. However, the prediction of MLR is (Min; 7.865, Max 59.844) significantly lesser than the actual (Min: 3.17, Max: 129.66) value, including RFR and ET.

SVR for predicting feedback request level (MAE: 0.384, RMSE: 0.550) and solving time (MAE: 8.065, RMSE: 14.763) has significantly performed worse than all other models. Furthermore, Table 5 shows SVR has an exploration issue on the prediction value (-0.081) for the feedback request level.

RFR for predicting feedback request level (MAE: 0.333, RMSE: 0.530) and solving time (MAE: 7.361, RMSE: 12.756), ET for predicting feedback request level (MAE: 0.335, RMSE: 0.530) and solving time (MAE: 7.461, RMSE: 12.894) had similar performance than MLR and SVR because ensemble models can be performed well on non-linear data. Moreover, the feedback request level's prediction output in Table 5 shows significantly similar with actual value (RFR: Min: 0.00 - Max: 2.780, ET: Min: 0.000 - Max: 3.000).

GBR for predicting feedback request level (MAE: 0.352, RMSE: 0.529) and solving time (MAE: 7.030, RMSE: 12.342) outperforms all other models. Moreover, for predicting solving time, the mean of GBR's prediction (17.327) is almost similar to the mean of MLR's prediction Mean (17.370).

In our analysis, every model has performed almost similarly shown in Table 4. However, for feedback request level, ensemble methods of RFR, ET; for solving time, GBR and MLR were found better than other models. One potential explanation is that the ensemble methods enrich the skills and generate more outcomes than expert linear models, which may also be a better model for the domain.

6. Discussion and Future Work

The study aims to explore a better method to estimate the mastery of math knowledge at the elementary level, which is taught by learning-by-units based on mastery-oriented goal principles. Our approach was to estimate students' mastery by predicting their feedback request level and solving time. The prediction model was applied to assess student's work on each task. The results revealed the plausibility of predicting students' feedback request level and solving time. Popular machine learning regression methods are used due to their broad adoption within the field of educational data mining and learning analytics and its' relative computational simplicity. The ensemble model of RFR and ET performed better than other models to predict future feedback request level; on the other hand, MLR and GBR outperform other models for the prediction of solving time.

The application of machine learning addresses how to build online learning systems that improve automatically through experience to assess student mastery and provide additional support. Prediction models used the log data of student's feedback request level and solving time of four questions to predict the student's feedback request level and solving time of the next question. Students are at various levels, have different skills and different learning abilities. Making high-quality predictions immediately available may help instructors, teaching assistants and intelligent learning systems identify groups of students who are wheel-spinning and need alternate forms of assistance (Slater & Baker, 2019). Without support, a student who is not making progress in an open and distance learning context is likely to drop out of the course (Tang, Xing, & Pei, 2018). If we can identify this struggle before it goes on for too long, we may be able to address the problem and help students get back on track. Researchers (Tuominen et al., 2020; Yu et al., 2018) have been done on recognizing the student's affective state and responding to it actively. The prediction model of this study can be modified for predicting students' affective state while the student's affective state data is collected.

Based on the result, the use of ML methods deserves more attention on the prediction of mastery or knowledge in future studies. Furthermore, this study encourages researchers interested in using ML methods and other cutting-edge knowledge tracing algorithms to predict mastery of the skill and make the path to move further direction (i.e., additional support or move on the next level), rather than just predicting performance within-data. Bälter (2018) described that OLI courses do not provide students with a path for moving on after mastering a skill, either through a forced pathway or by providing data to the student that he/she had mastered the skill. For adaptive behavior, accurate predictions matter most, while for actionable insights, the interpretability and the stability of parameter estimates supersede accuracy; open and distance learning learner models require both.

Overall, we make the following contributions: 1) our work makes the initial process of estimating students' mastery by predicting their feedback request level and solving time on the next question for developing an efficient adaptive mechanism. 2) We explored the robustness and effectiveness of the proposed models on mastery prediction tasks for using the dataset, which involved Math-Island online learning system, and 3) we explored the need to predict mastery in an online learning system autonomously. Our initial experiences have been very positive.

The present study has few limitations: This finding further emphasizes the necessity to use non-linear models to leverage historical data optimally. For solving time, the model has fit into the actual data very well; however, it seems more normalization tactics are needed for feedback request level. Furthermore, we used one learning unit's log data only; the result shows that more datasets and advanced deep learning and neural network models are needed, which is definitely included in our future research.

Although extending this study contains several possible directions to persevere. One such direction is incorporating these trained prediction models into a Math-Island tutoring system to estimate student mastery of skill and provide the amount of practice needed on time. Moreover, the interpretability of latent knowledge has not been fully explored, and further work is needed. For future work, we will not only explore how to design a mechanism, called, Mastery Prediction Model (MPM), to predict student mastery by adopting ML methods; but also investigate the following questions: 1) What is the minimum number of practice opportunities needed to achieve mastery of skill in mathematics learning; 2) Does the minimum opportunities vary by each domain in Math-Island; 3) After how many questions would it be suitable to predict students' mastery of skill, in order to make efficient adaptive learning context.

Acknowledgments

The authors would like to thank the Ministry of Science and Technology of the Republic of China, Taiwan, for financial support (MOST 109-2511-H-008 -011 -MY3), and Research Center for Science and Technology for Learning, National Central University, Taiwan. The authors would also like to thank all co-author for their assistance during this study.

References

Abidi, S. M. R., Hussain, M., Xu, Y., &Zhang, W. (2018). Prediction of confusion attempting algebra homework in an intelligent tutoring system through machine learning techniques for educational sustainable development. *Sustainability (Switzerland)*, *11*(1). https://doi.org/10.3390/su11010105

Ansems, E. L., Hanci, E., Ruijten, P. A. M., &IJsselsteijn, W. A. (2019). I focus on improvement: Effects of type of mastery feedback on motivational experiences. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11433 LNCS*, 213–224. https://doi.org/10.1007/978-3-030-17287-9_18

Asselman, A., Khaldi, M., &Aammou, S. (2020). Evaluating the impact of prior required scaffolding items on the improvement of student performance prediction. *Education and Information Technologies*, *25*(4), 3227–3249. https://doi.org/10.1007/s10639-019-10077-3

Bälter, O., Zimmaro, D., &Thille, C. (2018). Estimating the minimum number of opportunities needed for all students to achieve predicted mastery. *Smart Learning Environments*, *5*(1). https://doi.org/10.1186/s40561-018-0064-z

Beck, J. E., &Gong, Y. (2013). Wheel-spinning: Students who fail to master a skill. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7926 LNAI*, 431–440. https://doi.org/10.1007/978-3-642-39112-5-44

Caniëls, M. C. J., Chiocchio, F., &vanLoon, N. P. A. A. (2019). Collaboration in project teams: The role of mastery and performance climates. *International Journal of Project Management*, *37*(1), 1–13. https://doi.org/10.1016/j.ijproman.2018.09.006

Cen, H., Koedinger, K., &Junker, B. (2007). Is Over Practice Necessary ? – Improving Learning Efficiency with the Cognitive Tutor through Educational Data Mining. *Proceedings of the 13th International Conference on Artificial Intelligence in Education AIED 2007*, *158*, 511–518. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.142.7340&rep=%0Arep1&type=pdf

Chan, T.-W., Roschelle, J., Hsi, S., Kinshuk, Sharples, M., Brown, T., …Hoppe, U. (2006). One-To-One Technology-Enhanced Learning: an Opportunity for Global Research Collaboration. *Research and Practice in Technology Enhanced Learning*, *01*(01), 3–29. https://doi.org/10.1142/s1793206806000032

Corbett, A. T., &Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278. https://doi.org/10.1007/BF01099821

Gervet, T., Koedinger, K., Schneider, J., &Mitchell, T. (2020). When is Deep Learning the Best Approach to Knowledge Tracing? *Jedm*, *12*(3), 31–54.

Hussain, M., Zhu, W., Zhang, W., Abidi, S. M. R., &Ali, S. (2019). Using machine learning to predict student difficulties from learning session data. *Artificial Intelligence Review*, *52*(1), 381–407. https://doi.org/10.1007/s10462-018-9620-8

Imran, M., Latif, S., Mehmood, D., &Shah, M. S. (2019). Student academic performance prediction using supervised learning techniques. *International Journal of Emerging Technologies in Learning*, *14*(14), 92–104. https://doi.org/10.3991/ijet.v14i14.10310

Jin, C. (2020). MOOC student dropout prediction model based on learning behavior features and parameter optimization. *Interactive Learning Environments*, *0*(0), 1–19. https://doi.org/10.1080/10494820.2020.1802300

Lai, T. L., &Lin, H. F. (2015). A case study of the feedback design in a game-based learning for low achieving students. *Proceedings of the International Conference on E-Learning 2015, e-learning 2015 - Part of the Multi Conference on Computer Science and Information Systems 2015*, 213–214.

Lee, J. I., &Brunskill, E. (2012). The impact on individualizing student models on necessary practice opportunities. *Proceedings of the 5th International Conference on Educational Data Mining, EDM 2012*, 118–125.

Mao, Y., Lin, C., &Chi, M. (2018). Deep Learning vs. Bayesian Knowledge Tracing: Student Models for Interventions. *Journal of Educational Data Mining, 10*(2), 28–54. Retrieved from https://jedm.educationaldatamining.org/index.php/JEDM/article/view/318

Myneni, L. S., Narayanan, N. H., Rebello, S., Rouinfar, A., &Pumtambekar, S. (2013). An interactive and intelligent learning system for physics education. *IEEE Transactions on Learning Technologies*, *6*(3), 228–239. https://doi.org/10.1109/TLT.2013.26

Ogdanova, D. B., Tova, Y. U. A. K., &Vetaeva, K. N. E. T. C. (2014). Experts'' views consideration in assessing of the level of students'' knowledge in distance learning. *Вестник Уфимского Государственного Авиационного Технического Университета*, *18*(5 (66)), 102–104.

Pavlik, P. I., Cen, H., &Koedinger, K. . (2009). Performance Factors Analysis – A New Alternative to Knowledge Tracing. *Materials Letters, 212,* 531–538. Retrieved from http://dl.acm.org/citation.cfm?id=1659450.1659529

Pelánek, R. (2015). Metrics for Evaluation of Student Models. *JEDM - Journal of Educational Data Mining,* *7*(2), 1–19. Retrieved from http://www.educationaldatamining.org/JEDM/index.php/JEDM/article/view/JEDM087

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., &Sohl-Dickstein, J. (2015). Deep knowledge tracing. *Advances in Neural Information Processing Systems*, *2015-Janua*, 505–513.

Razzaq, R., Ostrow, K. S., &Heffernan, N. T. (2020). Effect of Immediate Feedback on Math Achievement at the High School Level. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-52240-7_48

Romero, M., Hernández, J. M., Juola, J. F., Casadevante, C., &Santacreu, J. (2020). Goal Orientation Test: An Objective Behavioral Test. *Psychological Reports*, *123*(4), 1425–1451. https://doi.org/10.1177/0033294119845847

Rong, S., &Bao-Wen, Z. (2018). The research of regression model in machine learning field. *MATEC Web of Conferences*, *176*, 8–11. https://doi.org/10.1051/matecconf/201817601033

Schnipke, D. L. (2002). Exploring Issues of Examinee Behavior: Insights Gained from Response-Time Analyses. *Computer-Based Testing: Building the Foundation for Future Assessments*, 237–266.

Slater, S., &Baker, R. (2019). Forecasting future student mastery. *Distance Education*, *40*(3), 380–394. https://doi.org/10.1080/01587919.2019.1632169

Sokkhey, P., &Okazaki, T. (2019). Comparative Study of Prediction Models for High School Student Performance in Mathematics. *IEIE Transactions on Smart Processing & Computing*, *8*(5), 394–404. https://doi.org/10.5573/ieiespc.2019.8.5.394

Sokkhey, P., &Okazaki, T. (2020). Hybrid machine learning algorithms for predicting academic performance. *International Journal of Advanced Computer Science and Applications*, *11*(1), 32–41. https://doi.org/10.14569/ijacsa.2020.0110104

Tang, H., Xing, W., &Pei, B. (2018). Exploring the temporal dimension of forum participation in MOOCs. *Distance Education*, *00*(00), 1–20. https://doi.org/10.1080/01587919.2018.1476841

Tuominen, H., Juntunen, H., &Niemivirta, M. (2020). Striving for Success but at What Cost? Subject-Specific Achievement Goal Orientation Profiles, Perceived Cost, and Academic Well-Being. *Frontiers in Psychology*, *11*(September), 1–18. https://doi.org/10.3389/fpsyg.2020.557445

Yeh, C. Y. C., Cheng, H. N. H., Chen, Z. H., Liao, C. C. Y., &Chan, T. W. (2019). Enhancing achievement and interest in mathematics learning through Math-Island. *Research and Practice in Technology Enhanced Learning, 14*(1). https://doi.org/10.1186/s41039-019-0100-9

Yu, L. C., Lee, C. W., Pan, H. I., Chou, C. Y., Chao, P. Y., Chen, Z. H., …Lai, K. R. (2018). Improving early prediction of academic failure using sentiment analysis on self-evaluated comments. *Journal of Computer Assisted Learning, 34*(4), 358–365. https://doi.org/10.1111/jcal.12247

Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Profiling Student Learning from Q&A Interactions in Online Discussion Forums

De Lin ONG, Kyong Jin SHIM* & Swapna GOTTIPATI
School of Computing and Information Systems, Singapore Management University, Singapore
*kjshim@smu.edu.sg

Abstract: The last two decades have witnessed an explosive growth in technology adoption in education. Proliferation of digital learning resources through Massive Open Online Courses (MOOCs) and social media platforms coupled with significantly lowered cost of learning has brought and is continuing to take education to every doorstep globally. In recent years, the use of asynchronous online discussion forums has become pervasive in tertiary education institutions. Online discussion forums are widely used for facilitating interactions both during the lesson time and beyond. Numerous prior studies have reported benefits of using online discussion forums including enhanced quality of learning, improved level of thinking beyond the classroom, collaborative knowledge building, and enhanced participation by shy or intimidated students. By monitoring and analyzing students' activities in online discussion forums, instructors can intervene and manage students' learning. For the instructor to employ appropriate intervention measures, both quantitative and qualitative analyses of students' participation are important. To mitigate the challenge of the sheer volume of conversation threads in online discussion forums, we present a text mining approach to profiling student learning based on Q&A interactions. Firstly, we perform text classification to categorize conversations into two categories: non-programming-related and programming-related. Secondly, from the programming-related conversation threads, our method categorizes students into four participation proficiency types based on their Q&A activities. Next, our method determines whether a student adopts more explicit or implicit expression behavior in Q&A activities. We evaluate our approach on the second-year computing course, Web Application Development II. Finally, we share the lessons learned in this teaching process.

Keywords: Profiling student learning, online discussion forum, Q&A interactions, slack

1. Introduction

Today, the Internet is ever present in the lives of many people globally. Proliferation of digital learning resources through Massive Open Online Courses (MOOCs) and social media platforms coupled with significantly lowered cost of learning is continuing to bring education to every doorstep. Leaders of educational institutions found online learning to be a critical component in their long-term strategic planning (Allen & Seaman, 2015). Online learning allows a wider range of individuals – such as full-time workers, caretakers, and mothers – to flexibly start and continue their learning journey (Ragusa & Crampton, 2017).

In recent years, the use of asynchronous online discussion forums has become pervasive in tertiary education institutions. Online discussion forums are widely used for facilitating interactions both during the lesson time and beyond. The traditional online discussion forum is typically situated within a Learning Management System (LMS). Today more than ever, as we live through the COVID-19 pandemic where virtual distance learning is commonplace, the benefits of online discussion forums are being revisited.

In online discussion forums, conversations can be neatly organized by course topics by using threads or channels. Digitally saved and viewed anytime anywhere, online discussion forums allow for learning much more flexibly compared to traditional modes of learning restricted to physical classrooms. Picciano's study (Picciano, 2002) found that the learner's participation in online discussion forums led to enhanced learning in terms of quantity and quality. Meyer's study reports that participation in online discussions is associated with improved level of thinking beyond the classroom

(Meyer, 2003). Another prior study reported a positive correlation between the learner's final grade and his engagement in online discussion forums (Bliuc et al., 2010).

Online discussion forums enhance the participation of shy or intimidated learners who would otherwise keep quiet in face-to-face lessons (Groeling, 1999; Al-Salman, 2009; Gerbic, 2010). Further, online discussion forums promote networking with other learners such that they can build knowledge collaboratively (Gilbert & Dabbagh, 2005). By monitoring and analyzing learners' activities in online discussion forums, instructors can intervene and manage students' learning (Stephens-Martinez, 2014; Jiang et al., 2015). Understanding the extent to which learners participate in the online discussion forum is important. Simply quantifying learners' participation such as the number of posts in online discussion forums does not adequately portray their contributions (Mazzolini & Maddison, 2007). For the instructor to employ appropriate intervention measures, qualitative analyses involving verification of each student's contribution are still deemed important. However, it can be quite challenging for the instructor to manually inspect the sheer volume of messages in an online discussion forum. One way to mitigate this problem is to employ an automated approach to the identification of relevant discussion forum threads. Further, giving instructors the ability to automatically profile learners based on learners' participation will allow instructors to quickly identify and narrow down to those learners that need most help at the moment.

Therefore, the goal of our paper is to present a text mining approach to profiling student learning based on Q&A interactions in Slack, a popular messaging application designed for teamwork. Originally built for businesses, by March 2020, over 3,000 colleges and universities had adopted Slack in their classrooms (Slack, 2020). Slack, Discord, Microsoft Teams, and Google Hangouts are similar platforms, and we choose Slack due to some free features compared to other tools (Tuhkala & Kärkkäinen, 2018). Firstly, we perform text classification to categorize conversations into two categories: non-programming-related and programming-related. Secondly, from the programming-related conversation threads, our method categorizes learners into four participation proficiency types based on their Q&A activities. Next, our method determines whether a learner adopts more explicit or implicit expression behavior in Q&A activities. We evaluate our approach on the second-year computing course, Web Application Development II. Finally, we share the lessons learned in this teaching process.

2. Background

2.1 Web Application Development Curriculum

In our university's Information Systems undergraduate program, all students must complete core courses including Web Application Development I (WAD I) and Web Application Development II (WAD II). WAD I focuses on server-side web development using HTML and PHP. WAD II is a follow-up course whose focus is on front-end web development, and it builds on top of WAD I by introducing students to HTML, CSS, Bootstrap and responsive UI, JavaScript, APIs, and Vue.js as part of the curriculum. In WAD II, 4–5-person group projects allow students to explore beyond-the-curriculum frameworks and tools such as ReactJS, AngularJS, Tailwind CSS, D3.js, Git, and other front-end frameworks/libraries as well as visualization and animation APIs. WAD II is a young course which was offered as a core course for the first time in the Fall 2020 semester.

2.2 Using Slack to facilitate Q&A in Web Application Development Courses

In early March 2020, all lessons at our university moved to fully online due to the COVID-19 pandemic. As a result, WAD I lessons and assessments abruptly moved to online via Zoom and our university's LMS with little preparation. At the time, only a few WAD I sections and faculty members had been using Slack to facilitate Q&A. While our lessons were still conducted fully on-campus, most students sought help from instructors and teaching assistants (TAs) via physical means by arranging face-to-face consultations on campus. Thus, while Slack served a good purpose for information dissemination, it was still not fully utilized as more students resorted to face-to-face Q&A and were not familiar with or comfortable with participating in online Q&A discussions.

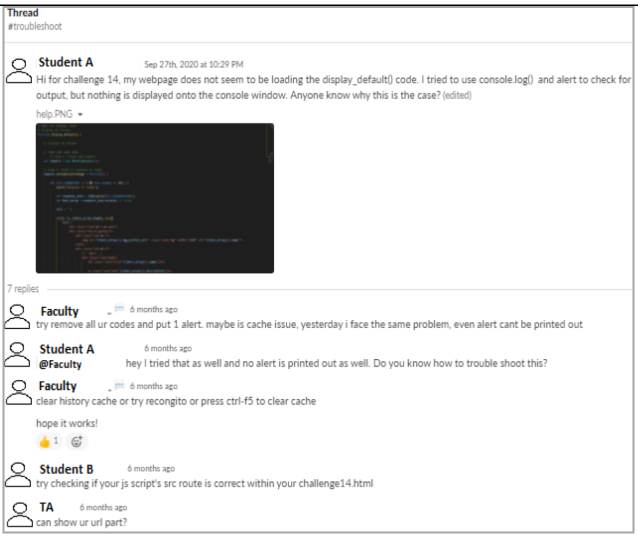
In the subsequent Fall semester, our university allowed a small percentage of courses to adopt hybrid classroom mode while a greater majority had to conduct all lessons fully online. Given the very hands-on nature of web development, WAD II instructors opted for the hybrid classroom option. In the hybrid classroom mode, students would take turns on a bi-weekly basis to alternate between online learning (work-from-home) and on-campus learning (in a designated classroom with appropriate social distancing measures in place) such that only about half of the class (~ 24 students per section) were allowed in a classroom each week. One week prior to the first lesson, only one-third of the students indicated in a survey that they would attend physical lessons. With a larger percentage of students learning-from-home while attending synchronous in-class lessons via Zoom, we determined that there would be a large gap in students' learning in terms of seeking help which would have been done in a much more convenient and efficient way by simply waiving a hand to ask questions or turning to peers sitting nearby in the classroom.

To overcome this challenge, some of the WAD II instructors adopted Slack to facilitate Q&A both inside-of-the-lesson and outside-of-the-lesson. Unlike the existing online discussion forum built into our university's LMS, Slack offers many useful and fun features suitable for programming-related courses and for the younger generation of students. For example, Slack offers a wide range of "stickers" (or emojis) that users can use to give "thumbs up" to show gratitude and "clapping" to congratulate and encourage others. As human body language and verbal tone may not fully translate in text messages, stickers or emojis serve as viable alternate means for communicating nuanced meaning (Tan, 2017). Further, Slack's "channels" provide an efficient way of organizing and searching through topic-based conversations. Lastly, Slack's API and data export feature (freely available for public channels) would allow for deep analyses of conversation threads in terms of identifying trending topics, profiling student learning based on Q&A activities and their interactions with other users, and development of automated analyses via programmatic means.

3. Methodology

3.1 Slack Terminology

Instructors can create Slack workspaces for free. The free version comes with maximum 10,000 messages limit, but there is no limit to the number of users that can join a workspace.



Thread
#troubleshoot

Student A Sep 27th, 2020 at 10:29 PM
Hi for challenge 14, my webpage does not seem to be loading the display_default() code. I tried to use console.log() and alert to check for output, but nothing is displayed onto the console window. Anyone know why this is the case? (edited)

help.PNG

7 replies

Faculty 6 months ago
try remove all ur codes and put 1 alert. maybe is cache issue, yesterday i face the same problem, even alert cant be printed out

Student A 6 months ago
@Faculty hey I tried that as well and no alert is printed out as well. Do you know how to trouble shoot this?

Faculty 6 months ago
clear history cache or try recongito or press ctrl-f5 to clear cache

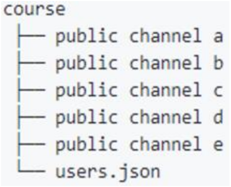
hope it works!

👍 1

Student B 6 months ago
try checking if your js script's src route is correct within your challenge14.html

TA 6 months ago
can show ur url part?

(Slack conversation thread in #troubleshoot channel)



```
course
├── public channel a
├── public channel b
├── public channel c
├── public channel d
├── public channel e
└── users.json
```

(Unzipped Slack workspace)

Figure 1. Slack Conversation Thread & Unzipped Slack Workspace from Data Export.

Inside a workspace, one or more 'channels' can be created. A 'channel' is equivalent to a discussion 'forum', and each channel is a separate space where multiple conversation 'threads' can occur. Figure 1 (left) shows an actual conversation thread from WAD II course where usernames are anonymized. The thread reads from top to bottom, and it was started by 'Student A' posting a question (message) about a JavaScript issue. Subsequently, three other users (Faculty, Student B, and TA) replied to Student A's message. Similar to popular social media platforms such as Facebook, Instagram and Twitter, Slack allows users to 'mention' other users using '@' (at) symbol. Mentioned users are then notified.

3.2 Dataset

At the start of the semester, instructors and TAs regularly checked all threads to determine which threads are of importance to the teaching team for intervention purposes. In particular, the WAD II course teaching team was interested in Q&A discussions on course-related topics. To automate the identification of programming-related threads, we downloaded the WAD II Slack workspace via Slack's export feature so that computer programs could perform analyses. The data was extracted as zip files via Slack's management console. Figure 1 (right) depicts an example of an unzipped workspace. The exported file contains all public channels as well as user data (in users.json file), and it does not contain Direct Messages (DMs) as DMs are private message exchanges between pairs of individuals. A total of 1,772 messages from threads were used for our analyses.

3.3 Identification of Programming-Related Q&A Messages

In this section, we describe a text classification method for identifying programming-related Q&A messages. The objective was to classify Q&A messages into either programming-related or non-programming-related. As shown in Figure 2, we first performed text preprocessing which included lowercasing all words, removing all the Slack's mentions, normalizing certain words, and expanding contractions and slangs. The resulting texts were then transformed into Word2Vec (Mikolov et al., 2013) vectors to train a classifier. Next, we labeled our dataset so that our model could learn to classify. Various models were experimented with the Word2Vec vectors and labels using Positive-Unlabeled Learning (Sansone et al., 2019). An evaluation of the various models was done against F1, Precision, and Recall. We chose the best model based on the highest F1 score, and the best model would be used to predict whether a message was programing related or not at a later stage. As there were no labels for identifying which messages were programming-related, we needed rules or heuristics for labeling them. As WAD II course focuses on front-end web development, we used certain rules to automatically label the dataset to determine if the message was programming-related or not.

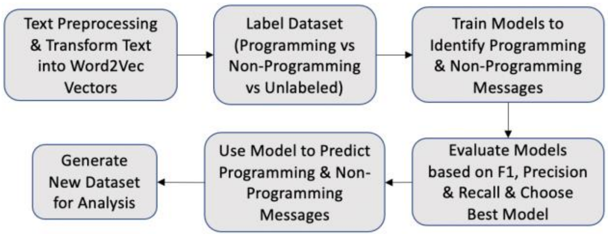

Text Preprocessing & Transform Text into Word2Vec Vectors
Label Dataset (Programming vs Non-Programming vs Unlabeled)
Train Models to Identify Programming & Non-Programming Messages
Evaluate Models based on F1, Precision & Recall & Choose Best Model
Use Model to Predict Programming & Non-Programming Messages
Generate New Dataset for Analysis

Figure 2. Summarized Workflow for Text Classification.

At the initial stage of labeling, we used three labels. The three labels were: 1) programming related, 2) non-programming-related, 3) un-labeled. With Positive-Unlabeled learning, we converted unlabeled text into either programming-related or non-programming-related. The rationale behind having a label called 'unlabeled' was because we used certain rules to label our dataset. However, our rules were not comprehensive, and it might not have picked up certain characteristics of web development-related terms such as HTML, CSS, and JavaScript keywords. Hence, instead of simply passing these texts as non-programming-related, we used Positive-Unlabeled Learning so that the classifier could learn the hidden characteristics of the unlabeled dataset by itself.

```
def label_message(text):
    If text contains <= 1 word:
       Return non-programming-related label
    Else:
       If the results for HTML tags are valid:
          Return programming-related label
       Else If there are results for JavaScript:
          Return programming-related label
       Else If there are results for CSS with ";":
          Return programming-related label
       Else If there are results for CSS with {}:
          Return programming-related label
       Else If the results for HTML tags are invalid:
          Return non-programming-related label
       Else:
          Return unlabeled
```

Figure 3. Pseudocode for Labeling Messages.

```
def find_html(text):
   Find HTML tags using <> within text
   If there are any HTML tags:
       Check if links and symbols are valid html tags or are programming-related
   Do another check to remove other stop words
   Return result
def find_javascript(text):
    Find Javascript using ";" and return result in a list
def find_css_colons(text):
    Find CSS using ":" and return result in a list
def find_css_brackets(text):
    Find CSS using "{" and "}" and return result in a list
```

Figure 4. Pseudocode for checking if a Message Contains HTML, CSS, JavaScript.

We first checked if there were more than one word found in the message text (Figure 3). If the text was only one-word long, the text was non-programming-related as HTML, CSS & JavaScript could not be done with a single word. Next, we performed a series of checks to see if HTML, CSS, and JavaScript code was found in the text (Figure 4). We checked if there were any HTML tags in the message using regular expressions. As URLs were also embedded in tags, we had to include certain URLs as stop words. If there were any HTML tags, we labeled them as programming-related. If the text contained ";", it was likely JavaScript. Hence, we labeled it as programming-related. We performed similar checks subsequently for CSS by checking the existence of ";", "{", and "}".

After performing all checks for HTML, CSS, and JavaScript, if the text was not yet labeled, we double checked to see if the text had previously been identified as containing HTML syntax but filtered away as invalid. In that case, we assigned non-programming-related label. Having non-programming-related labels was important as these labels would serve as reliable negatives for our classifiers to be trained using Positive-Unlabeled Learning.

To determine the classifier to be used for our research, we experimented with k-nearest neighbors (k-NN), Support Vector Machine (SVM), and XGBoost algorithms. We only used Word2Vec representation as our features to train the classifiers. Classifiers were trained using Positive-Unlabeled Learning which assumed two-class classification, but there were no labeled negative examples for training. The training data was only a small set of labeled positive examples and a large set of unlabeled examples.

There are mainly four methods in Positive-Unlabeled Learning: Rocchio, Naive Bayesian Classifier, Spy technique, and 1-DNF method. The Spy technique inserts positive labels (spies) into the unlabeled dataset. A prior study (Liu et al., 2003) reported the results of using two-step strategy for learning a classifier from positive and unlabeled data with a detailed evaluation of all the combinations of methods. Their results indicated that the models trained using the Spy technique achieved fairly high F scores (above ~0.8 macro-averaged F-score). Therefore, in our study, we decided to use the Spy technique. Classification performance results are shown in Table 1. SVM achieved the best result in terms of F1 and recall while XGBoost achieved the highest precision. Thus, for the final predictive model, we chose the SVM classifier with the highest F1 score.

Table 1. *Classifier Performance*

Classifier	F1	Recall	Precision
K-Neighbors	0.859	0.810	0.914
SVM	0.897	0.886	0.909
XGBoost	0.872	0.823	0.929

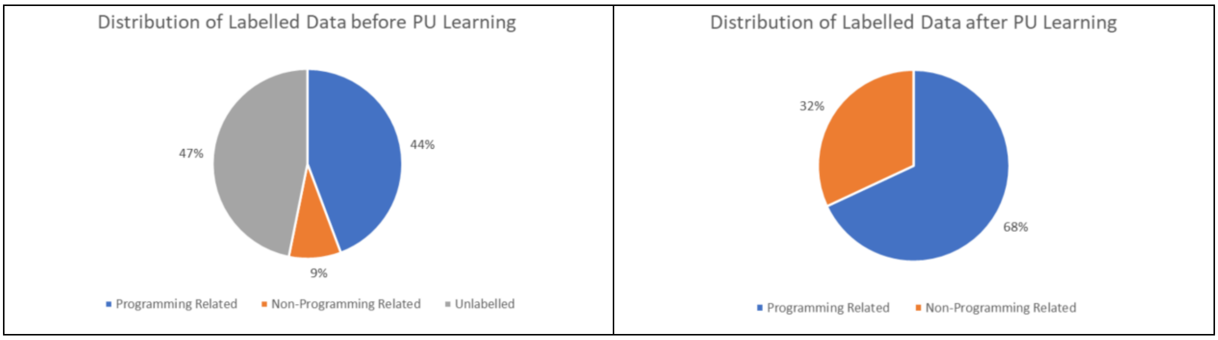


Figure 5. Distribution of Labeled Data before & after PU Learning.

As shown in Figure 5 (left), using our rule-based (heuristic) approach of labeling, 44% of the messages were labeled as programming-related, and 47% were labeled as non-programming-related. Further, after performing Positive-Unlabeled Learning, we could capture even more programming-related messages. As shown in Figure 5 (right), 68% of the messages were now labeled as programming-related and 32% labeled as non-programming related. Next, we performed an error analysis by calculating model confidence.

Table 2. *Top 5 Messages Incorrectly Predicted to be Non-Programming-Related by PU Learning*

Message	Model Confidence
when you use float: right, the space beside the container with float:right becomes "empty" and the html code that comes after will be "moved up" to fill that "empty" space. when you add float: right to #side-bar, the code that comes after in the html code which is the main-footer is "moved up" and takes the "empty" space to the left of your side-bar which is empty when you did float:right. you should create "empty" space to the right of the main-content so that the side-bar container can "move up"	92.9%
might want to try refer to this <https://www.w3schools.com/css/css_link.asp.You> could have possibly just clicked the link the style applied may not have been reflected	89.2%
The clear property specifies how elements should behave when they bump into each other on the page. 1.left: the left side of the element will not touch any other element within the same containing element. 2.right: the right side of the element will not touch any other element within the same containing element. 3.both:neither side of the element will touch any other element within the same containing element. 4.none: the element can touch either side	87.1%
can try the following to push the list style inwards to your content list-style-position: inside;	86.1%
Try the code in the following sequence In your chrome browser: 1. Directly visit/load post.php 2. Go to form.html, fill out the form and press SUBMIT (which will lead you to post.php) 3. Go to form2.html, fill out the form and press SUBMIT (which will lead you to post.php) 4. Go to form_empty.html, fill out the form and press SUBMIT (which will lead you to post.php)	81.9%

Table 3. *Top 5 Messages Correctly Predicted to be Programming-Related by PU Learning*

Message	Model Confidence
Hi guys, Please remember to add the <h1> tag with you own name. There are some websites which missed out this portion.	51.0%
hi anyone know why such behaviour is displaying? I am currently using font-size:150% for the button which by right should display 1.5x bigger than the rest of the font and it looks fine when rendered as a website but when rendered in mobile version, the font shrinks and looks smaller in comparison to the rest of the text (Web Vs Mobile-Iphone 5)	42.7%
prof, for $status in add, if we didnt prompt the user for values of status, can we	41.8%

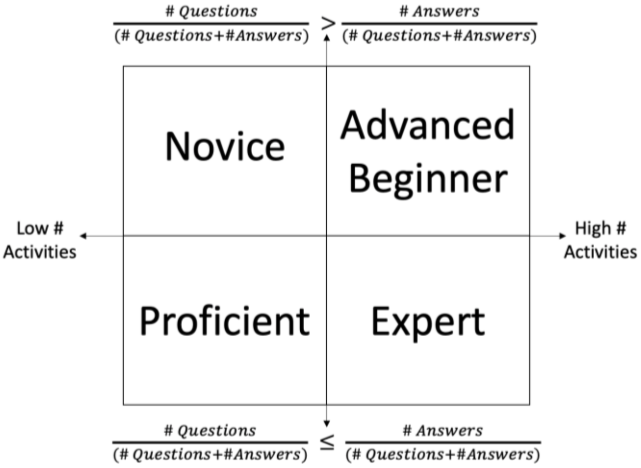
just use variable of $status = 'A' from the start and not use :status?	
```echo "Hello, my name is $person";```	41.2%
navigate to the 2 folders, inside look for api/config/database.php inside line 9 ```private $password = "";``` change password to empty if your on wamp. or if you manually changed your own mysql password	38.0%

Table 2 shows the top five messages that were previously labeled as programming-related (by rule-based approach), and our Positive-Unlabeled Learning model incorrectly predicted them to be non-programming-related. We observe that the model confidence is in the higher range. Table 3 shows the top five messages that were previously labeled as programming-related (by rule-based approach), and our Positive-Unlabeled Learning model correctly predicted them to be programming-related. The model exhibits a lower overall confidence in this case.

This indicates that our rule-based approach might not be perfectly suitable for labeling messages despite high F1, precision, and recall scores. The rules we used might be more suited to identify code blocks, but the messages found in Slack are in the form of Q&A "discussions" where users not only post code examples but also write their thoughts and explanations in English. This can result in texts where plain English is mixed with code snippets which may not necessarily be syntactically correct (e.g. missing closing semicolon). Therefore, the heuristics in our rule-based approach must be flexible enough to account for such cases. One way to mitigate this problem might be to leverage human judges to assess and label the initial pool of Q&A messages, and the heuristics can be further refined based on verified human-labelled messages.

3.4 Q&A Participation Proficiency Types

From the programming-related messages, we quantified the number of questions and the number of answered made by each student. Note that we only included 'explicit' expressions for the purposes of determining proficiency types *(and this is further elaborated in Section 3.5).*



$\frac{\#\,Questions}{(\#\,Questions+\#Answers)} > \frac{\#\,Answers}{(\#\,Questions+\#Answers)}$
Novice
Advanced Beginner
Low # Activities
High # Activities
Proficient
Expert
$\frac{\#\,Questions}{(\#\,Questions+\#Answers)} \leq \frac{\#\,Answers}{(\#\,Questions+\#Answers)}$

Figure 6. Q&A Participation Proficiency Types.

We categorized students into four Q&A participation proficiency types based on the number of questions and the number of answers they contributed in the Slack channels of the WAD II course. Figure 6 shows a quadrant of the four proficiency types. The *novice* are users who are not active and tend to ask more questions than to provide answers. The *advanced beginner* are users who are active and tend to ask more questions than to provide answers. The *proficient* are users who are not active and tend to provide more answers than to ask questions. The *expert* are users who are active and tend to provide more answers than to ask questions.

3.5 Q&A Participation Expression Types

As described in Section 2.2, Slack has great support for a wide range of stickers to convey meaning. We define 'explicit' behavior to be direct textual verbalization of questions or answers. For example, if a student posts a piece of code and writes *"I don't understand why this code"* in the message field, this is considered an 'explicit' expression. On the other hand, if the student presses a "confused" sticker to *indicate* that he is confused (instead of writing out the text), we consider this an implicit expression.

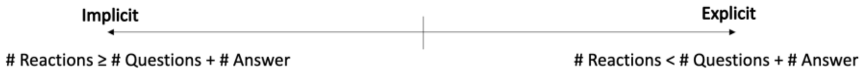

Implicit
Explicit
# Reactions ≥ # Questions + # Answer
# Reactions < # Questions + # Answer

Figure 7. Implicit vs. Explicit Expression Types.

As shown in Figure 7, if the number of stickers is more than or equal to the number of questions and answers, we categorize the student as exhibiting more of an 'implicit' expression behavior. On the other hand, if the number of stickers is less than the number of questions and answers, the student is considered to be exhibiting more of an 'explicit' expression behavior where he tends to verbalize questions and answers without using stickers.

4. Findings & Discussions

4.1 Q&A Participation Proficiency Types & Expression Types

From the programming-related messages, using the Q&A Participation Proficiency quadrant (Figure 6) and Expression Types scale (Figure 7), we were able to profile student learning as shown in Figure 8.

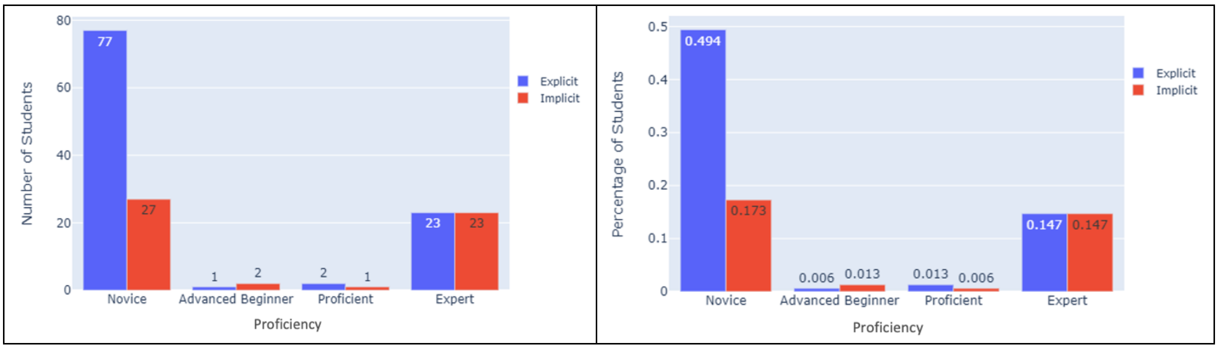


Figure 8. Student Learning Profiles (Proficiency Type & Interaction Type).

As shown in Figure 8 (right), over 66% of the students are categorized as "Novice" and over 29% of the students are categorized as "Expert" in programming-related Q&A discussions in Slack. Only a small number of students belong to "Advanced Beginner" and "Proficient" categories. Overall, the distribution of Q&A participation proficiency types appears to be un-even across the four proficiency types. There are two important findings that are valuable to WAD II course instructors.

Finding (A) ***"Novice" Students*****:** They are not active and tend to ask more questions than to provide answers. There are two possible scenarios behind this: 1) "Novice" students can seek help independently (via Internet search or other means). Thus, they do not have any or too many questions to post or 2) they need help but are shy or intimidated to post questions publicly. Figure 8 shows the results from our end-term analysis where we profiled all students at the end of the semester over all Slack Q&A threads created up until that point. To employ appropriate and timely intervention measures, it is recommended that instructors perform regular profiling of student learning. By profiling student learning on a more frequent basis throughout the semester, instructors can verify whether students fall into the first scenario or the second one by giving frequent pop-quizzes or frequent knowledge checks.

1) In the first scenario, pop-quizzes will help students verify their own knowledge of course topics. It will also indicate to instructors which of the "Novice" students need help. Instructors can encourage the "Novice" students who perform well on pop-quizzes to answer other students' questions (in which case, they could bump up to "Proficient" or even "Expert" levels).
2) In the second scenario, instructors and TAs must intervene and offer help with course content. If students are too shy to post questions on their own, instructors can encourage

them to check Slack channels regularly and give stickers (‘implicit’ expression type) to questions posted by other students that they resonate with. Although it is an ‘implicit’ form of expression, from the sticker types, instructors can reasonably infer the students’ intention and be able to tell whether certain students need help. In the exported Slack dataset, all stickers are shown as emoji texts (e.g. :confused:, :help:, :im_lost) that can be easily parsed and analyzed to detect sentiment.

Finding (B) *“Expert” Students*: the presence of over 29% of “Expert” students is very promising as they actively partake in answering other students’ questions and helping maintain WAD II Slack workspace a vibrant discussion space. In terms of expression type, they are equally split into “Explicit” and “Implicit”. Manual inspection of Slack threads reveals that the “Expert” students actively give stickers to other students in a form of encouragement (e.g. claps, cheers, thumbs up). Additionally, Q&A threads where “Expert” students who exhibit “Explicit” expression type point us to the next direction in our research. As mentioned in Section 2.1, WAD II had its first run as a core course in Fall 2020, and it repeats every Fall semester. From the Fall 2020 Slack Q&A threads, we can mine Frequently-Asked-Questions and corresponding answers provided by the “Expert” and “Explicit” students to build a knowledge repository which can be re-used in the future. Such a knowledge repository will help instructors better understand which topics are confusing to students and thus need more revision and hands-on exercises. It will also greatly help new students quickly find answers to programming-related questions.

4.2 Q&A Participation and Performance

We used Pearson correlation coefficient scores between Q&A activities in Slack and student course grade to analyze the impact of online discussion forum participation on grades (Table 4). There is no significant evidence suggesting any correlation. In our analysis, we observe that while Q&A activities are highly skewed, student grades are normally distributed. This observation explains the lack of correlations between Q&A activities and student grades.

Table 4. *Correlations between Q&A Activities and Student Performance*

	# Answers	# Questions	# Reactions	# Messages	Total # Activities
Pearson Coefficient r	0.207	0.191	0.152	0.221	0.242
P-Value	0.0124	0.0215	0.0684	0.0075	0.0034

5. Limitations & Future Work

This research work is based on the case study of one faculty member to profile student learning. While we do not aim to generalize our findings, the technical steps of text classification involving heuristics and machine learning as well as categorization of students into Q&A participation proficiency types and expression types suggest a viable “automated” way of analyzing large volumes of conversation threads. As mentioned in Section 3.3, the existing rule-based approach can be further improved to identify programming-related messages. In our analysis, the percentage of “Novice” students was quite high at 66%. In our case study, we were not able to conclusively tell whether these students belong to the first or second scenario as described in Section 4.1. During the case study period, due to the COVID-19 pandemic, it was logistically challenging to plan pop-quizzes and fully automate student profiling. We plan to conduct another larger scale case study now with a fully automated student profiling system.

6. Conclusion

In this paper, we present a text mining approach to profiling student learning based on Q&A interactions in online discussion forums. Firstly, we perform text classification to categorize conversations into two categories: non-programming-related and programming-related. Secondly, from the programming-related conversation threads, our method categorizes students into four participation proficiency types (Novice, Advanced Beginner, Proficient, Expert) based on their Q&A activities. Proficiency types are determined based on the number and ratio of questions and answers posted by students. Next, our method determines whether a student adopts more explicit or implicit expression behavior in Q&A activities. Students resorting to stickers (emojis) more than words are considered to be exhibiting "implicit" expression type. Students that tend to use more English words to verbalize their thoughts and emotions than stickers are considered to be exhibiting "explicit" expression type. We evaluated our approach on a second-year undergraduate course, Web Application Development II.

Acknowledgements

We would like to thank all teaching team members and students who participated in this study.

References

Allen, I. E., & Seaman, J. (2015). Grade Level. Tracking Online Education in the United States. Babson Survey Research Group and Quahog Research Group, LLC.

Al-Salman, S. M. (2009). The role of the asynchronous discussion forum in online communication. *Journal of Instruction Delivery Systems*, 23(2), 8–13.

Bliuc, A., Goodyear, P., Ellis, R. (2010). Blended learning in higher education: How students perceive integration of face-to-face and online learning experiences in a foreign policy course. *Research and Development in Higher Education: Reshaping Higher Education*, Melbourne, July 6–9, vol. 33, pp. 73–81.

Gilbert, P. K., Dabbagh, N. (2005). How to structure online discussions for meaningful discourse: a case study. *Br J Educ Technol.*, 36(1):5–18.

Gerbic, P. (2010). Getting the blend right in new learning environments: A complementary approach to online discussions. *Education and Information Technologies*,15, 125–137.

Groeling, T. (1999). Virtual Discussion: Web-based Discussion Forums in Political Science. *Paper presented at the 1999 National convention of the American Political Science Association,* Atlanta, Georgia.

Hrastinski, S. (2008). What is online learner participation? A literature review. *Computers & Education*, 51(4), 1755–1765.

Jiang, Z., Zhang, Y., Liu, C., Li, X. (2015). Influence analysis by heterogeneous network in MOOC forums: what can we discover? *Int Educ Data Min Soc*, 242–249.

Liu, B., Dai, Y., Li, X., Lee, W. S., & Yu, P. S. (2003). Building text classifiers using positive and unlabeled examples. *Third IEEE International Conference on Data Mining*, 179-186.

Mazzolini, M., & Maddison, S. (2007). When to jump in: The role of the instructor in online discussion forums. *Computers & Education.*, 49. 193-213. 10.1016/j.compedu.2005.06.011.

Meyer, K. (2003). Face-to-face versus threaded discussions: The role of time and higher-order thinking. *Journal of Asynchronous Learning Networks*, 7(3), 55–65.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ICLR*.

Picciano, A. G. (2002). Beyond student perceptions: Issues of interaction, presence, and performance in an online course. *Journal of Asynchronous Learning Networks*, 6(1).

Ragusa, A. T., & Crampton, A. (2017). Online learning: Cheap degrees or educational pluralization? *British Journal of Educational Technology*, 48(6), 1208–1216.

Sansone, E., Natale, F. D., & Zhou, Z. (2019). Efficient Training for Positive Unlabeled Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41, 2584-2598.

Slack. (2020, March 16). Distance learning thrives in Slack. https://slack.com/intl/en-sg/blog/collaboration/distance-learning-in-slack

Stephens-Martinez, K., Hearst, M. A., Fox, A. (2014). Monitoring moocs: which information sources do instructors value? *Proceedings of the first ACM conference on Learning@ scale conference*, p. 79–88.

Tan, R. L. (2017, July 23). How emojis have changed the way people communicate. https://www.straitstimes.com/lifestyle/how-emojis-have-changed-the-way-people-communicate

Tuhkala, A., & Kärkkäinen, T. (2018). Using Slack for computer-mediated communication to support higher education students' peer interactions during Master's thesis seminar. *Education and Information Technologies.*, 23. 10.1007/s10639-018-9722-6.

Wise, A. F., Marbouti, F., Hsiao, Y., Hausknecht, S. (2012). A survey of factors contributing to learners 'listening' behaviors in asynchronous online discussions. *J Educ Comput Res*, 47(4):461–80.

SLP: A Multi-Dimensional and Consecutive Dataset from K-12 Education

Yu LU[a], Yang PIAN[a*], Ziding SHEN[b], Penghe CHEN[a], Xiaoqing LI[a]
[a]*Advanced Innovation Center for Future Education, Beijing Normal University, China*
[b]*University of California, Los Angeles, USA*
* bianyang@mail.bnu.edu.cn

Abstract: Learning is a complicated process jointly influenced by multiple factors, such as learner's personal characteristics, family background and school environment. However, the existing public datasets in K-12 education domain seldom fully cover the heterogeneous dimensions, which greatly hinders the research on fully analyzing and understanding the learners and their learning process. In this work, we report a dataset that includes the learners' demographic information, psychometric intelligence scores as well as their family-school background information. Furthermore, the dataset records the learners' academic performance data on 8 different subjects in 3 consecutive years. This multi-dimensional dataset from K-12 education can be a valuable information source for learning analytics and would benefit the cross- disciplinary research in education on a broader canvas. The dataset has been publicly available for the research purpose at https://aic-fe.bnu.edu.cn/en/data/index.html.

Keywords: Educational dataset, learning analytics, K-12 education

1. Introduction

Driven by the fast advancements of online learning platforms and educational data collection techniques, learner modelling and learning analytics (Ferguson, 2012) have recently attracted much attention from both academia and industry. Proper collection and analysis of educational data have then become crucial, as they provide opportunities to fully understand the human learning process and precisely identify its underlying key factors, which may ultimately help improve human's learning experience, efficiency and efficacy.

Specifically, the Massive Open Online Course (MOOC) platform and Intelligent Tutoring System (ITS) (Polson & Richardson, 2013) have significantly facilitated the collection of learner's online assessment data and their interaction with the corresponding virtual learning environment. For example, the ASSISTments dataset (Selent, Patikorn & Heffernan, 2016) includes the assessment data of secondary school students and the log data recording their interactions with an ITS for math study. The KDD Cup dataset (Qiu et al., 2016) contains the MOOC learners' assessment and interaction data on 40 different online courses. The EdNet dataset (Choi et al., 2020) gathers learner's online interaction in 4 different levels of abstractions from a multi-platform service. The Open University Learning Analytics Dataset (OULAD) (Kuzilek, Hlosta & Zdrahal, 2017) introduces learner's demographic information, including gender, geographic region and education background. The Programme for International Student Assessment (PISA) dataset (OECD, 2015) covers the participants' assessment data, and school-family information, which mainly used the one-time test (performed every three years on 15-year-old students) without continuous tracking on the participants. A public data repository, called PSLC DataShop (Stamper, 2010), has been established to host the education-related datasets, particularly for the learning interaction data.

However, there seldom exists dataset that provides multiple heterogeneous dimensions and could be directly employed for interdisciplinary research purpose. Past literature has already demonstrated that psychometric intelligence is a strong predictor of student's scholastic success, which can be measured by the Intelligence Quotient (IQ) test (Downey et al., 2014; Eccles & Harold, 1993). Several external factors have then been proven to influence IQ score, in which family and parental factors might directly impact it, especially for K-12 learners (Haimovitz & Dweck, 2016; Melby et al., 2008). For example, students who are from a higher socioeconomic background, a smaller-sized family or with a better educated parent, might be more intelligent and perform better at school (Von Stumm & Plomin, 2015). Lack of such heterogeneous dimensions could hamper the further exploration on learning analytics, and thus would possibly prevent researchers from discovering the essential laws of education.

To fully cover all these heterogeneous dimensions, our dataset collected from an online learning platform called smart learning partner (SLP), intentionally recorded the learner's data from the five dimensions above to provide a wealthy content for learning analytics and educational data mining. It thus has two unique characteristics:

- It explicitly covers the data from five different dimensions, namely student demographic information, psychometric intelligence information, academic performance information, family information and school information;
- It automatically captures the learner's academic performance data during their three-year study (mainly from 7th grade to 9th grade) on 8 different subjects, namely Math, Physics, Chemistry, Biology, History, Chinese, Geography and English.

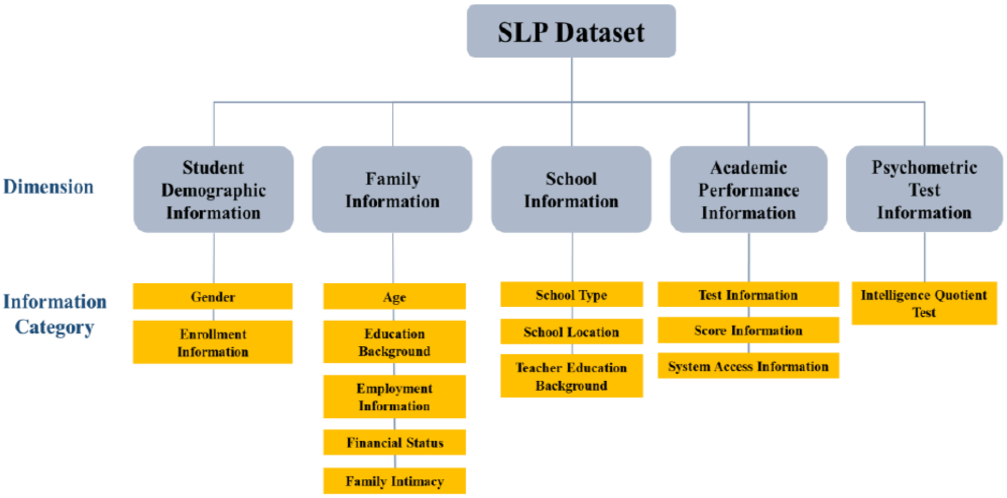


Figure 1. Overall Structure of SLP Dataset.

Figure 1 gives the overall structure of the SLP dataset. Briefly speaking, the SLP dataset are collected from 4830 students in 32 local secondary schools. The entire data collection process has lasted approximately three years (from November 2016 to June 2019) by leveraging on an online learning platform, called SLP, which currently serves more than 140,000 secondary school students. The students employ this platform to conduct regular unit tests and term tests since their first year in secondary school (i.e., 7th grade) to the last year (i.e., 9th grade). Up to June 2019, a total of 832 unit tests and 54 term tests on 9 different subjects have been designed and implemented on the platform. Specifically, each unit test is associated with a knowledge concept in a given subject, while each term test normally covers the knowledge concepts that have been learned during the entire semester. In addition, 14,027 learning resources (mainly in the form of micro-lecture, i.e., a short video in 5 to 10 minutes) would be automatically recommended to students based on their unit test or term test results.

The SLP dataset could be used in different studies, such as predicting student's academic performance in a fine-grained manner (Chen et al., 2018), or evaluating the influences of pertinent factors (e.g., intelligence or family factors) on different subject learning. Our aim is to encourage researchers from diverse fields to engage in understanding and modelling learners in K-12 education.

2. Data Collection Workflow

Figure 2 depicts the data collection process from different sources: the student basic information and the school information were directly acquired from the local education bureau, mainly consisting of the student demographic and enrolment information as well as the corresponding school information. Such information was also utilized to create and validate the individual user's account on the online learning platform, which guaranteed all the platform users were the authenticated local students. When a student logged into the platform for the first time, he/she was required to complete an online survey to collect his/her family information, where the questionnaire contains the questions about family intimacy, parent education background and other pertinent family information. The first-time login and the data collection process were usually guided by local teachers during class time, and it thus, to a certain extent, ensured the high quality of the collected family information data.

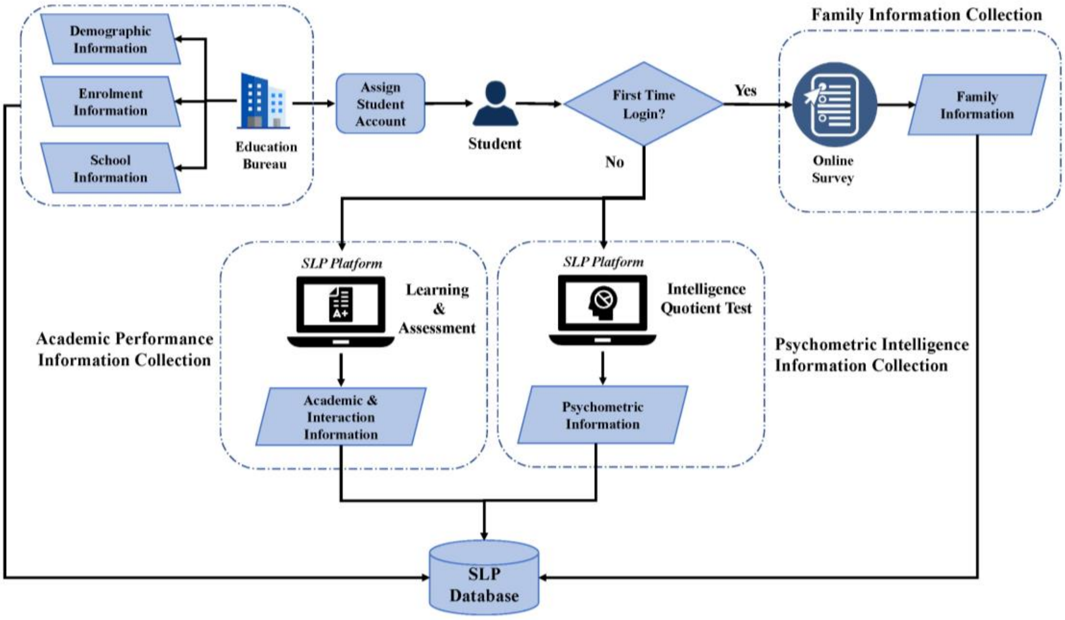


Figure 2. SLP Data Collection and its General Workflow.

2.1 Assessment Data Collection

After the first-time login and the online family survey accomplished, students were asked to regularly use the platform to conduct the assessment tests, and to freely choose the personalized learning resources on different subjects. All of the test results, including both unit tests and term tests, and the important interaction log data are automatically recorded and stored by the platform. A unit test normally contains 9 questions on the same knowledge concept and mostly in the form of multiple-choice questions (MCQ) or short-answer free text questions. When students were doing a unit test, the platform recorded students' access time. Most of the unit tests could be auto-scored by the system. On the other hand, the term test typically contains more questions on multiple knowledge concepts in a similar form, and the access time was also recorded. The SLP dataset includes all the above information, as well as the name of the knowledge concept tested by each question and the name of the corresponding subject.

2.2 Psychometric Intelligence Data Collection

On the SLP platform, students were offered with multiple psychometric tests and could choose to take any of these based on their preference. In the current SLP dataset, 1161 students participated and responded to one part of the IQ test online. The test was based on the classical Raven's Progressive Matrices Test (Raven & Court, 1938) with minor modifications to fit with the cultural and linguistic context. Specifically, students answered a 40-item ($\alpha = .80$) question set. Scores were then computed based on norms, and the total score was 140 with 70 as the passing line. A high score would indicate a mastery of reasoning ability, demonstrating the student's ability to deduce, summarize, and exchange the pattern with others solely based on the provided information.

2.3 Data Usage and Privacy Issues

Before the data collection process, all the participants, including the students, their parents and schools, have been explicitly informed that the collected data on the platform would be directly used for the research purpose and shared publicly. All the parents were given the option to opt out of the data sharing before they signed the data usage agreement. After the data was submitted, we also carefully notified each parent via both Email and SMS, and removed the ones who felt uncomfortable to be involved in the SLP dataset. To further protect the privacy of the opt-in participants, we had irreversibly anonymized their identity, and randomly sampled a subset of the participants.

3. Data Description

The SLP dataset is available in the form of multiple CSV files (value is separated by comma, and the first line in each table shows the column name), and can be accessed from the website[1]. Figure 3 illustrates tables and corresponding data fields in the SLP dataset, which would be elaborated as below.

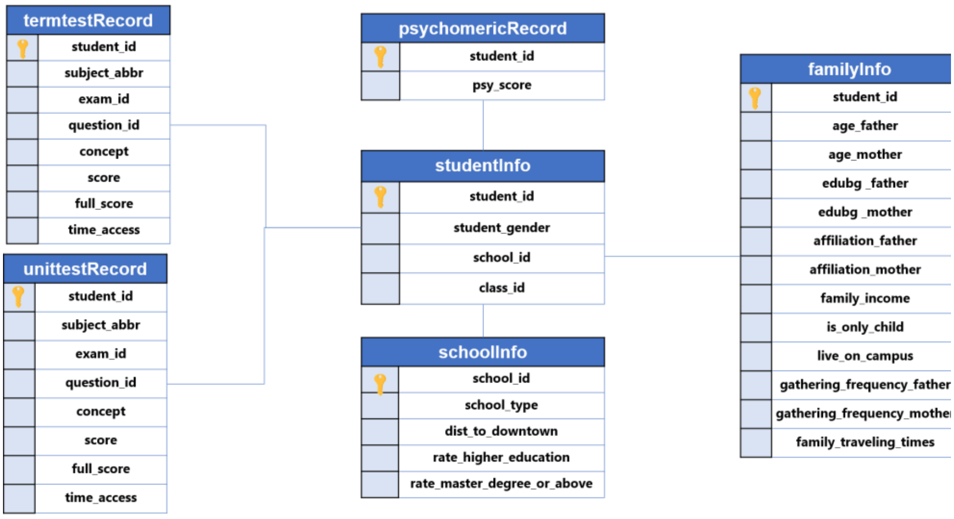


Figure 3. Tables and Data Fields of SLP Dataset.

- **Academic Performance Information**: The two file folders, named “termtestRecord” and “unittestRecord”, contain multiple tables with the identical table structure as above, which provide the student academic performance information. Specifically, these two folders collect the term test and unit test data respectively. In each folder, each sub-table contains individuals’ subject score information. For example, the sub-table “termtestRecord-BIO” inside the folder “termtestRecord” stores the term test results of Biology, which currently consists of 500,562 rows. Similarly, the sub-table “unittestRecord-MATH” inside the folder “unittestRecord” keeps the unit test results of Math, which currently consists of 57,244 rows. Note that multiple concepts may be associated to one question, which would be separated by semicolon, and all the associated concepts are explicitly shown in the table (e.g., “line segment” or “intersecting lines” in math questions, “digestive system” or “urinary system” in biology questions).
- **Psychometric Test Information**: Table *psychomericRecord* contains the student’s psychometric intelligence score, and it consists of 1161 rows. Note the SLP dataset only includes participants who agreed to take and share their psychometric test results.
- **Family Information**: Table *familyInfo* contains student’s family information, including parents’ age, education background, employment information, financial status and family intimacy (e.g., how often the student meets with father or mother). The table consists of 3189 rows, and only includes the students who completed the family information survey and confirmed the truth of it.
- **School Information**: Table *schoolInfo* contains the school’s demographic information, including the school type, location information and teacher’s education background. It consists of 32 rows, and for teachers’ education background, 3 school’s data are missing (marked as n.a. in the table).
- **Student Demographic Information**: Table *studentInfo* contains the basic demographic information of students, including their gender, school and class ID. The table consists of 4830 rows, and all the school and class information are anonymized.

[1] https://aic-fe.bnu.edu.cn/en/data/index.html

4. Technical Validation

The SLP dataset contains learners' information from multiple dimensions, ranging from academic achievement, psychometric intelligence, to student's family and school information. We thus tentatively provide some illustrations and simple analysis results to validate the quality and value of the dataset, while deeper and broader studies could be further conducted.

We first looked at the SLP data from the perspective focusing on the relationship between family factors and intelligence. Figure 4 compares the mean of IQ test score within different family information categories. Each bar and the number above it represent the mean psychometric score in one group within one category. The analysis showed that there was a significant difference of psychometric intelligence score between students from one-child family (M = 87.40, SD = 9.99) and students from multiple-child family (M = 85.81, SD = 9.35). In addition, there was a significant positive correlation between students' psychometric test score and family's annual income (r = .07, n = 1161, p <.05), showing that a higher socioeconomic family background might be beneficial on children's intelligence development. More interestingly, when we analyzed the relationship between parents' education level and students' intelligence, only mother's education level was found to be significantly correlated with the psychometric test scores in a positive way (r = .12, n = 1161, p <.05), indicating that mothers with higher education background might have kids who also score higher in the intelligence test.

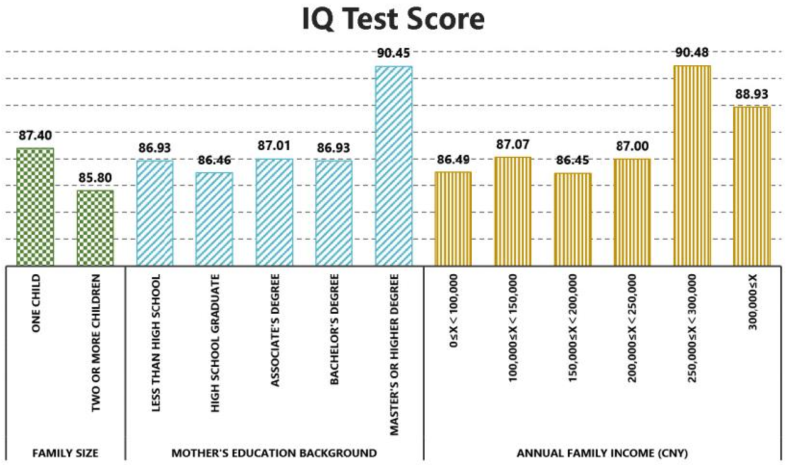


Figure 4. Mean Psychometric Test Scores Grouped by three Categories of Family Information.

Secondly, we focused on the relationship between school and academic performance. After removing the schools with data size smaller than 100 students, we portraited a scatter plot with different points representing each school's average term test score and the distance from school to downtown area, as shown in Figure 5. The line represents the significant negative correlation (r = -.87, n = 11, p <.05). This analysis result might reveal a unique pattern in China that most of the large and high-quality primary and middle schools aggregate around the center of the city.

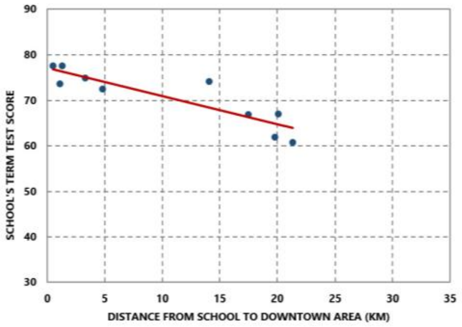


Figure 5. Scatter Plot of School's Term Test Scores and Its Distance to Downtown.

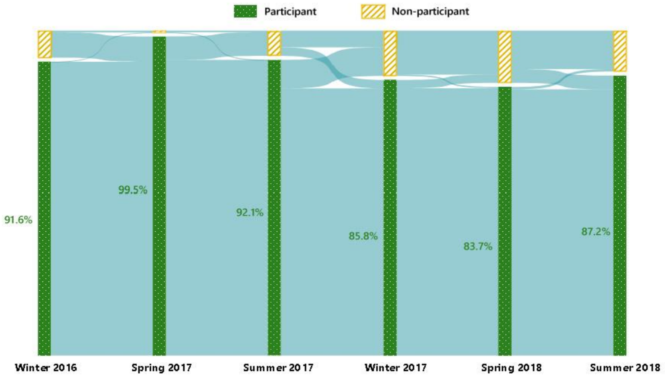


Figure 6. Sankey Diagram Describing the Term Test Participants' Flow Over Time.

Lastly, we validated the continuous collection of the academic performance data from the same student population across three years. Figure 6 shows the Sankey diagram that portrays how the student population has changed from 2016 to 2018 in 6 different term tests, where the width of the flow is proportional to the flow quantity. As could be seen from Figure 6, a majority of the students took all 6 term tests: for example, around 91.6% and 99.5% students took the term test in winter of 2016 and spring of 2017, whereas more than 90% of students took both tests. The diagram clearly validated the consistency of the academic performance data in terms of participants.

5. Conclusion

In this paper, we present an education-oriented 3-year consecutive dataset from k-12 learners. This dataset consists of multi-dimensional data including the learner's demographic information, academic performance, psychometric intelligence scores, and family-school background. We tentatively conduct some preliminary studies and provide some simple analysis results to validate the quality and value of the dataset. We believe the SLP dataset would provide valuable information and foster the cross-disciplinary research for learning analytics on a broader canvas.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No. 62077006, 61807003), and the Fundamental Research Funds for the Central Universities.

References

Chen, P., Lu, Y., Zheng, V.W., Pian, Y. (2018). Prerequisite-driven deep knowledge tracing. In *2018 IEEE International Conference on Data Mining (ICDM)* (pp. 39-48). IEEE.

Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., ... & Heo, J. (2020). Ednet: A large-scale hierarchical dataset in education. In *International Conference on Artificial Intelligence in Education* (pp. 69-73). Springer, Cham.

Downey, L. A., Lomas, J., Billings, C., Hansen, K., & Stough, C. (2014). Scholastic success: Fluid intelligence, personality, and emotional intelligence. *Canadian Journal of School Psychology*, *29*(1), 40-53.

Eccles, J. S., & Harold, R. D. (1993). Parent-school involvement during the early adolescent years. *Teachers college record*, *94*, 568-568.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5-6), 304-317.

Haimovitz, K., & Dweck, C. S. (2016). Parents' views of failure predict children's fixed and growth intelligence mind-sets. *Psychological science*, *27*(6), 859-869.

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open university learning analytics dataset. *Scientific data*, *4*(1), 1-8.

Melby, J. N., Conger, R. D., Fang, S. A., Wickrama, K. A. S., & Conger, K. J. (2008). Adolescent family experiences and educational attainment during early adulthood. *Developmental psychology*, *44*(6), 1519.

OECD. (2015). Programme for international student assessment (PISA). Retrieved April 30, 2021, from http://www.oecd.org/pisa/publications/.

Polson, M. C., & Richardson, J. J. (Eds.). (2013). *Foundations of Intelligent Tutoring Systems*. Psychology Press.

Qiu, J., Tang, J., Liu, T. X., Gong, J., Zhang, C., Zhang, Q., & Xue, Y. (2016). Modeling and predicting learning behavior in MOOCs. In *Proceedings of the ninth ACM international conference on web search and data mining* (pp. 93-102).

Raven, J. C., & Court, J. H. (1938). Raven's progressive matrices. Los Angeles, CA: Western Psychological Services.

Selent, D., Patikorn, T., & Heffernan, N. (2016). Assistments dataset from multiple randomized controlled experiments. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (pp. 181-184).

Stamper, J., Koedinger, K., d Baker, R. S., Skogsholm, A., Leber, B., Rankin, J., & Demi, S. (2010). PSLC DataShop: A data analysis service for the learning science community. In *International Conference on Intelligent Tutoring Systems* (pp. 455-455). Springer, Berlin, Heidelberg.

Von Stumm, S., & Plomin, R. (2015). Socioeconomic status and the growth of intelligence from infancy through adolescence. *Intelligence*, *48*, 30-36.

Learning Analytics Dashboard Prototype for Implicit Feedback from Metacognitive Prompt Responses

May Kristine Jonson CARLON & Jeffrey S. CROSS*
School of Environment and Society, Tokyo Institute of Technology, Japan
*cross.j.aa@m.titech.ac.jp

Abstract: Online learning can be challenging to learners as they need to have autonomous learning skills to succeed, and to instructors as direct observation and real-time communication with learners are limited. Learning analytics dashboards have been used to assist the learners in developing autonomous learning skills and the instructors in keeping track of the learners' progress. However, there is little information on systems supporting both learners and instructors in online learning environments. This paper builds on our previous work developing learners' metacognitive skills through open response prompts by using the learner inputs to create a dashboard that uncovers implicit feedback such as sentiments, misconceptions, and shallow learning. The instructor can consult the dashboard on-demand, and the input is from metacognitive prompts that only the individual learners see. Hence, the instructor can provide timely interventions based on inputs from learners who otherwise would not voice their concerns in more public channels such as discussion forums.

Keywords: Topic modeling, sentiment analysis, text similarity, metacognitive prompting, learning management system, learning analytics

1. Introduction

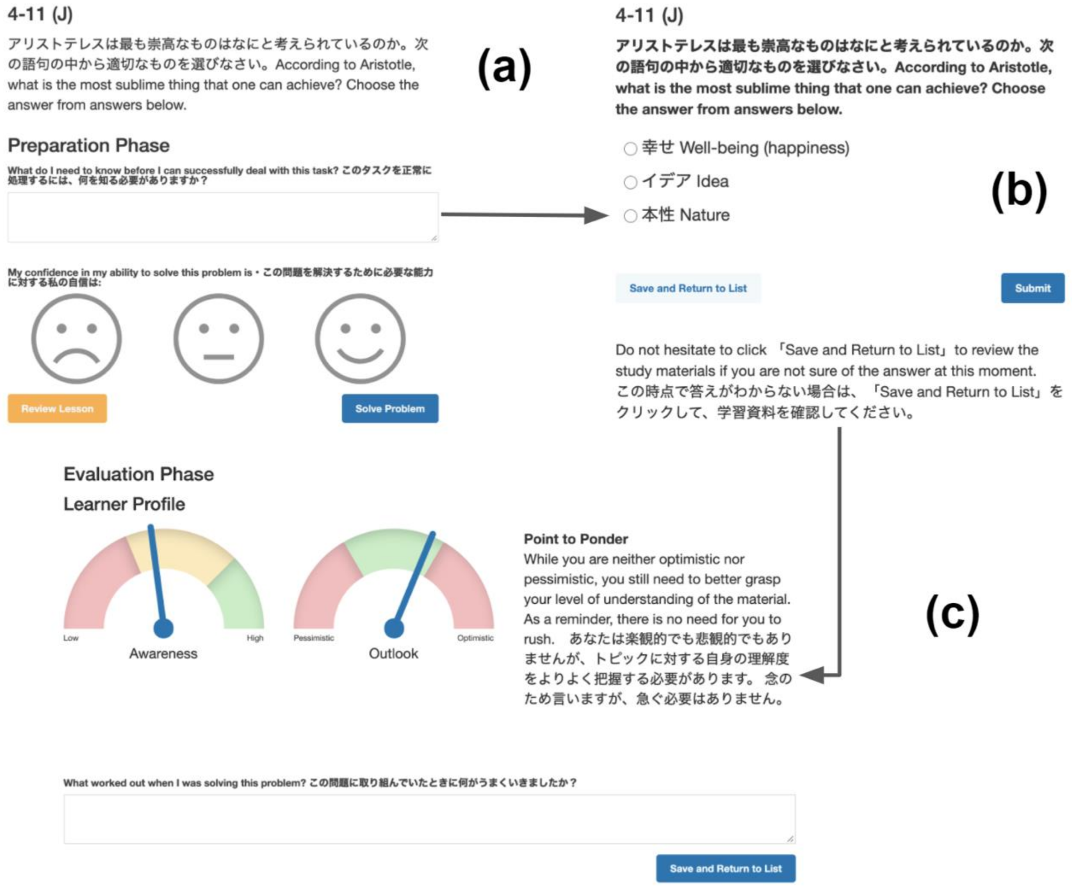
In a traditional classroom, learners have multiple avenues of interaction with the instructor available to them. When the learner needs something, they can raise their hand during class or approach the instructor after the class finishes or during office hours. Likewise, the instructor can gauge the learners' engagement by observing what is happening inside the classroom and with follow-up interactions outside of class. Even if the classroom interactions are minimal, there are still opportunities for the instructors and learners to exchange feedback during several activities, including homework, projects, and exam evaluation. In an online classroom, such communication can be more difficult. For those in synchronous classes, such as through teleconferencing, in most cases, the learners are just muted, and the instructor is not even sure if the learners are present. In other forms of the online classroom, such as in asynchronous classes like massive open online courses (MOOCs), learners and instructors can interact in the discussion forums. However, research has shown that in some cases, only about three percent of learners post on MOOC discussion forums (Onah et al., 2014). Another possibility for interaction is when learners provide feedback during course surveys. However, these surveys are frequently done at the start and the end of the course. Hence, there is not much opportunity for the instructor to adjust their teaching based on learner feedback. Additionally, most online classrooms are self-paced, so it is hard for them to see how learners are faring overall. An added complication is that some learners lack self-directed learning skills, which are essential to succeeding in online learning environments (Zhu et al., 2020).

An important thing to consider moving forward is that online learning has its strengths and weaknesses, and it is likely here to stay. The COVID-19 pandemic situation remains uncertain despite vaccination roll-out worldwide. We also have learned that some learners benefit from online platforms (Reich, 2021). For example, some learners have jobs with schedules conflicting with their classes, or some learners have disabilities that are better accommodated in an online learning platform. So, it is only fair to keep online classrooms or hybrid forms as an option for everyone who might need them now that we know that it is plausible for all levels of learning.

2. Related Work

Learner feedback on instructional style and materials is critical for improving teaching quality for future learners (Feistauer and Richter, 2017). Periodic feedback is also needed to prevent learners from losing motivation (Nicol and Macfarlane-Dick, 2006). However, student evaluation surveys, which are the most common source of instructor feedback, can suffer from a lack of validity, bias, and lack of utility (Marsh and Roche, 1997). An alternative source of feedback that recently has been gaining attention is learning analytics dashboards. While learning analytics dashboards can benefit learners in developing self-directed learning skills (Sedrakyan et al., 2020), they might prevent instructors from exercising creativity in instructional planning and decision making (Brown, 2020).

Self-directed learning can be handled separately by introducing metacognitive tutors in the learning environment (Gama, 2005). An example of this is the Personalized Online Adaptive Learning System (POALS) Metacognitive Tutor we developed, shown in Figure 1. This tool was designed to work either as a stand-alone web-based application or as an add-on to learning management systems (LMS) such as Canvas or Moodle. The POALS Metacognitive Tutor divides a learning activity into three phases - Preparation, Problem-Solving, and Evaluation – to allow the learners to reflect metacognitively using open-response prompts. The POALS Metacognitive Tutor has been instrumental in developing the learners' ability to regulate their knowledge, which is vital for self-directed learning (Carlon and Cross, 2021). Details about the utility of the POALS Metacognitive Tutor (e.g., how metacognition is measured, expected inputs to the prompts, and others) in metacognitive instruction is discussed in our other works.



4-11 (J)

アリストテレスは最も崇高なものはなにと考えられているのか。次の語句の中から適切なものを選びなさい。According to Aristotle, what is the most sublime thing that one can achieve? Choose the answer from answers below.

Preparation Phase

What do I need to know before I can successfully deal with this task? このタスクを正常に処理するには、何を知る必要がありますか？

My confidence in my ability to solve this problem is・この問題を解決するために必要な能力に対する私の自信は:

4-11 (J)

アリストテレスは最も崇高なものはなにと考えられているのか。次の語句の中から適切なものを選びなさい。According to Aristotle, what is the most sublime thing that one can achieve? Choose the answer from answers below.

○ 幸せ Well-being (happiness)
○ イデア Idea
○ 本性 Nature

Do not hesitate to click 「Save and Return to List」 to review the study materials if you are not sure of the answer at this moment. この時点で答えがわからない場合は、「Save and Return to List」をクリックして、学習資料を確認してください。

Evaluation Phase

Learner Profile

Point to Ponder
While you are neither optimistic nor pessimistic, you still need to better grasp your level of understanding of the material. As a reminder, there is no need for you to rush. あなたは楽観的でも悲観的でもありませんが、トピックに対する自身の理解度をよりよく把握する必要があります。念のため言いますが、急ぐ必要はありません。

What worked out when I was solving this problem? この問題に取り組んでいたときに何がうまくいきましたか？

Figure 1. POALS Metacognitive Tutor with the Preparation (a), Problem-Solving (b), and Evaluation (c) screens in Japanese and English.

There are existing works on learning analytics dashboard for online learning environments. Some of these quantitative measures from learner performance but can be limited in their ability to enable qualitative assessment such as classroom climate. Those that allow qualitative evaluation through natural language processing (NLP) approaches typically rely on discussion forums and course surveys whose data may not be representative of everyone due to the problems mentioned earlier. By using the POALS Metacognitive Tutor which is a private channel between the educator and the learners and can possibly be made mandatory, we can have an analytics data source deriving inputs from everyone. Metacognitive instruction may also introduce new dynamics in the online classroom that is yet to be explored for analytics. This paper demonstrates how the POALS Metacognitive Tutor can be used as input for a learning analytics dashboard by constructing a prototype of the NLP-based POALS Analytics Dashboard. A user study was also conducted to evaluate the perceived usefulness and usability of the POALS Analytics Dashboard.

3. The POALS Analytics Dashboard

The POALS Analytics Dashboard is based on the human-in-the-loop (HITL) design or interfaces that require human interaction. While HITL was previously attributed to increased problems in computer systems due to human errors in the past (Cranor, 2008), HITL is increasingly recognized to build fairer AI systems (Zanzotto, 2019). For the POALS Analytics Dashboard, the information is presented to the instructor in a digestible manner, and further actions were left for the instructor to decide (i.e., no recommendations provided). Since decision to action is still left to the instructors, POALS Analytics Dashboard does not aim to be a substitute for classroom interaction but as a supporting tool. The dashboard will be made up of three visualizations, each discussed separately, to uncover feedback that the learners might not provide to the instructors in an online environment.

The POALS Metacognitive Tutor described earlier was deployed in an electrical engineering course offered to first-year undergraduate students at Tokyo Institute of Technology (Tokyo Tech) over two academic quarters: from September 2020 to February 2021. Each quarter was made up of seven weeks: three weeks were delivered synchronously through teleconferencing and four weeks asynchronously as a small private online course (SPOC) on the LMS edX Edge. The resulting visualizations shown below are derived chiefly from metacognitive prompt responses completed by 29 students.

3.1 Sentiment Analysis

The online instructor might be interested to know how happy their learners are in their class. Subtle, elusive, unverbalized emotions are the basis of thought, meaning, and language, affecting perception and, eventually, cognition (Kanazawa, 2020). To address this, we conducted sentiment analysis on learner metacognitive prompt responses. Figure 2 (left) is a composite visualization of learner sentiments. The column graph shows the average absolute sentiment polarity score from the prompt responses. Its color changes to red when the sentiment is negative, orange when neutral, and green when positive. The ratio of learners with negative, neutral, or positive sentiments is shown through the half-donut chart.

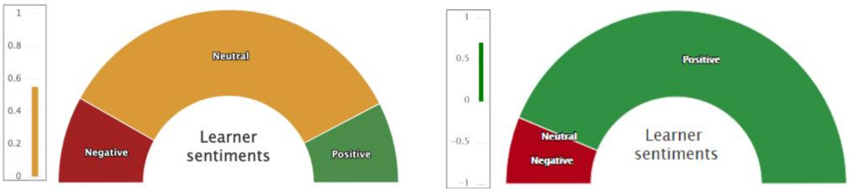


Figure 2. Ideal (left) and Resulting (right) Sentiment Analysis Visualization.

The learner responses in both the preparation and evaluation prompts were used as inputs for the sentiment analysis. All inputted text was presumed to be Japanese. The Bidirectional Encoder Representations from Transformers (BERT) was used as the pre-trained system for tokenizing, specifically bert-base-japanese-whole-word-masking. The bert-base-japanese-sentiment classifier was used as the pre-trained model, which labels the input as either positive or negative. In our case, it is crucial to see the percentages of learners who have polar sentiments; hence we need to establish a neutral range that is not predefined in our sentiment classifier choice. For both negative and positive cases, when the probability was less than 0.25, the label was changed to neutral. The cut-off is heuristically set and must be further investigated. These labels (positive, negative, neutral) corresponded to the green, red, and orange colors in the half-donut chart. The probabilities were multiplied by -1 when the original label is negative, and the values are summarized to get the value for the column graph.

Figure 2 (right) shows the result from the electrical engineering course, and it is noticeable that there are not many neutral responses. When the sentiment classifier was investigated, it was observed that even a very neutral sentence like "I live in Tokyo" gives a very positive score. A possible solution is to build a sentiment classifier specific to the task, but this requires extensive data collection that may not be easily scaled.

3.2 Topic Modeling

Another thing that the online instructor might think is, had the class been in-person, what topics will the learners be discussing? A word cloud like in Figure 3 (left) where the most prominent word in each topic tackled by the learners is shown and size is the probability assigned to the topic can be useful in answering this question.

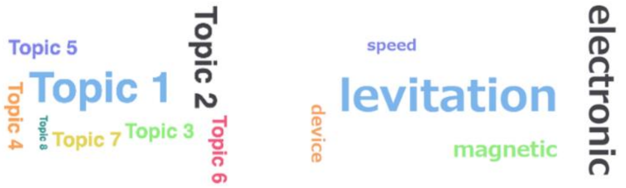


Figure 3. Ideal (left) and Resulting (Right) Topic Modeling Visualization.

To get the topic information, we conducted topic modeling on the metacognitive prompt responses using Latent Dirichlet Allocation (LDA), where observations (e.g., words in responses) are used to explain latent or unobserved groups (i.e., topics) by looking where the observations are present (Blei et al., 2003). The presence matrix, called tf-idf, was constructed by matching up the word frequency (tf: term frequency) with the responses the said word appears (idf: inverse document frequency).

Figure 3 (right) is the result from the electrical engineering class after translation to English. While knowing the topics the learners are thinking of is interesting, learning about these topics can also help the instructor uncover misconceptions. To elaborate, if a topic from a particular module comes up, it may be intuitive to think that the class average for the said module may be high since enough learners have thought carefully about it. If an emerging topic ends up coming from a low-scoring module, it may be worth investigating if there is a misconception that the learners are repeating. Misconception is a prevalent problem in education, may it be in chemistry (Uce and Ceyhan, 2019) or any other field.

3.3 Similarity Network

Finally, the instructor may be interested to know which topics the learners can remember well. Surface learning is the phenomenon where a learner picks up just enough knowledge to pass tests (Dolmans et al., 2016). The choice between surface learning or deep learning relies not just on the learners' motivation but also on the accompanying learning activities. The instructor must see if surface learning is widespread in their class, indicating concerns on how the learning activities are constructed instead of several individual motivations. The instructor may detect surface learning through a network graph, like the one shown in Figure 4 (left). Nodes represent modules, and related modules are connected; the weight of the lines expresses the degree of relatedness. Hovering a node shows the average learners' score in the exercises for the said module on a scale of 0 to 1. By inspecting the average learner scores on related modules, the instructor may detect surface learning.

The corresponding texts must first be converted into numerical representations to measure similarities between modules. The doc2vec algorithm available in the gensim library was used for vectorization. The resulting node weights then serve as the vector representation of the words. The vectors derived for each module might have different magnitudes (e.g., some modules having more words). The cosine similarity, which measures the angle between two vectors, was selected instead of the more popular Euclidean distance to reduce the effect of magnitudes.

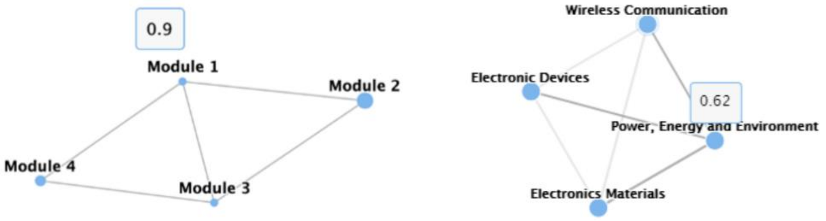


Figure 4. Ideal (left) and Resulting (Right) Similarity Network Visualization with Learners' Score.

Figure 4 (right) shows the result for the electrical engineering course. All similarity values are between 0.42 and 0.51. In this figure, we decided to make the nodes connected if their similarity score is above 0.4 due to proximity of similarity scores, so all nodes ended up being connected. This is a challenge with the proposed similarity network: what is the proper cut-off for similarity? We need to conduct similar experiments to have a better idea.

4. User Study

The POALS Analytics Dashboard was introduced to nine educators working in secondary schools, professional training, after-school support, and higher education (undergraduate and graduate) from Japan, the Philippines, United States, and Finland. They have experience in face-to-face and hybrid formats, and one has experience in a fully online format. They answered a questionnaire made up of four parts: a written interview to inquire about their experiences engaging with learners and their opinions about recent educational trends; an introduction to POALS Metacognitive Tutor and Analytics Dashboard; a Likert scale based on the Technology Acceptance Model (TAM) with a rating from 1 – Strongly Disagree to 5 – Strongly Agree (Davis, 1991); and a free-response form for further feedback.

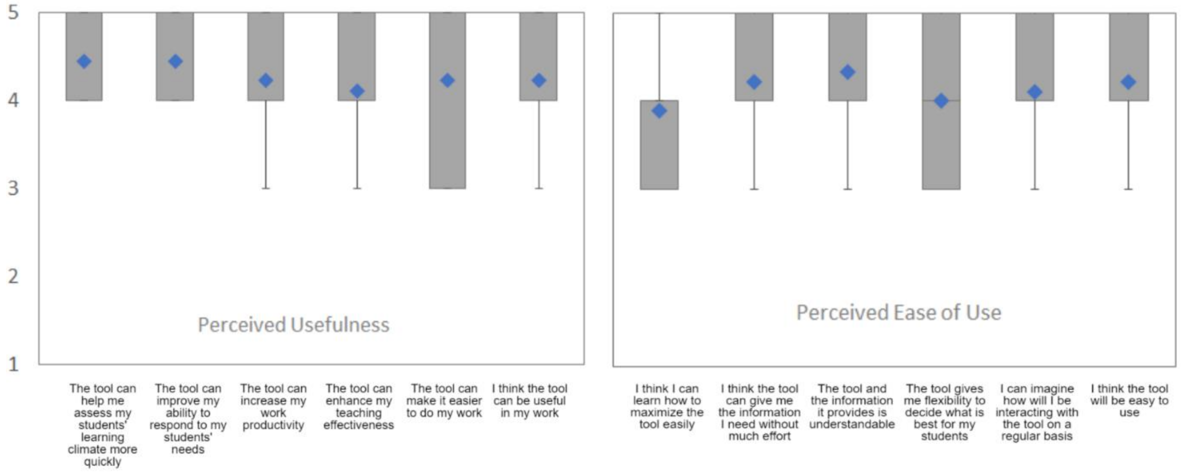


Figure 5. Boxplot of Modified TAM Results with Means Illustrated as Blue Diamonds.

Figure 5 shows the boxplots of the modified TAM responses. All the respondents agreed that the POALS Analytics Dashboard can help them respond to their learners' unvoiced needs and assess learner progress. Most of them had shown in their open question responses a keen interest with sentiment analysis as knowing the learners' feelings is a challenge in online learning environments. The only item that had a mean below 4 – Agree is the perception of maximizing the use of the tool. The respondents should be given more time to use the tool in their classes to understand this better.

5. Summary and Future Work

The POALS Analytics Dashboard is a learning analytics dashboard grounded on instructor intuition and learning theories. It provides details that can be used for detecting poor sentiments, misconceptions, and surface learning while leaving the judgment whether interventions should be made or not to the instructor. User study indicates a positive outlook towards the concept; thus, it is worthwhile to develop further the POALS Analytics Dashboard (e.g., multiple language support, more interactivity, etc.).

Acknowledgments

This research was supported by the Tokyo Tech Electrical Engineering Literacy teaching team consisting of Professors Akira Chiba, Takayuki Iwasaki, and Khanh Tran Gia. The Japan Society for the Promotion of Science (JSPS) supported this work via the Grants-in-Aid for Scientific Research (Kakenhi) Grant Number JP20H01719. Furthermore, the authors are indebted to the Center for Innovative Teaching and Learning for use the open edX LMS on the TokyoTechX edge.edx.org server.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993-1022.

Brown, M. (2020). Seeing students at scale: How faculty in large lecture courses act upon learning analytics dashboard data. *Teaching in Higher Education*, *25*(4), 384-400.

Cranor, L. F. (2008, April). A framework for reasoning about the human in the loop. In *Proceedings of the 1st Conference on Usability, Psychology, and Security* (pp. 1-15).

Carlon, M. K. J. & Cross, J. S. (2021, March). Open response prompts in an online metacognitive tutor. In *Program of the JSET Spring Conference 2021* (pp. 565-566).

Davis, F. D. "Perceived usefulness, perceived ease of use, and user acceptance of information technology." *MIS Quarterly* (1989): 319-340.

Dolmans, D. H., Loyens, S. M., Marcq, H., & Gijbels, D. (2016). Deep and surface learning in problem-based learning: a review of the literature. *Advances in Health Sciences Education, 21*(5), 1087-1112.

Gama, C. A. (2005). *Integrating metacognition instruction in interactive learning environments* (Doctoral dissertation, University of Sussex).

Feistauer, D., & Richter, T. (2017). How reliable are students' evaluations of teaching quality? A variance components approach. *Assessment & Evaluation in Higher Education*, *42*(8), 1263-1279.

Kanazawa, Y. (2020). Emotion is "deeper" than cognition: theoretical underpinnings and multidisciplinary lignes de faits to the Emotion-Involved Processing Hypothesis (EIPH). *国際学研究= Journal of International Studies*, *9*(1), 185-206.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: the critical issues of validity, bias, and utility. *American Psychologist*, *52*(11), 1187.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199-218.

Onah, D. F., Sinclair, J. E., & Boyatt, R. (2014, November). Exploring the use of MOOC discussion forums. In *Proceedings of London International Conference on Education* (pp. 1-4).

Reich, J. (2021). Ed tech's failure during the pandemic, and what comes after. *Phi Delta Kappan*, *102*(6), 20-24.

Sedrakyan, G., Malmberg, J., Verbert, K., Järvelä, S., & Kirschner, P. A. (2020). Linking learning behavior analytics and learning science concepts: Designing a learning analytics dashboard for feedback to support learning regulation. *Computers in Human Behavior*, *107*, 105512.

Üce, M., & Ceyhan, İ. (2019). Misconception in chemistry education and practices to eliminate them: literature analysis. *Journal of Education and Training Studies*, *7*(3), 202-208.

Zanzotto, F. M. (2019). Human-in-the-loop Artificial Intelligence. *Journal of Artificial Intelligence Research*, *64,* 243-252.

Zhu, M., Bonk, C. J., & Doo, M. Y. (2020). Self-directed learning in MOOCs: exploring the relationships among motivation, self-monitoring, and self-management. *Educational Technology Research and Development*, 1-21.

Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Analysis of the Answering Processes in Split-Paper Testing to Promote Instruction

Shin UENO*, Yuuki TERUI, Ryuichiro IMAMURA, Yasushi KUNO & Hironori EGI
[a]*Department of Informatics, The University of Electro-Communications Tokyo, Japan*
*u2130010@edu.cc.uec.ac.jp

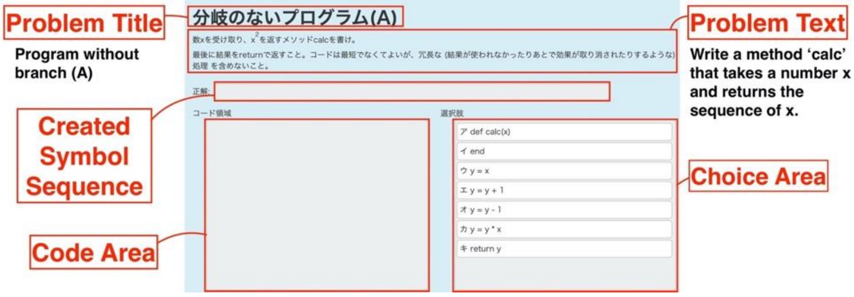
Abstract: This present study investigates the novice programmers using state transition diagrams, to help teachers and teaching assistants (TAs) in their instruction by visualizing the answering process of students. The data for this study are examination results in Ruby programming. Three state transition diagrams, correct answerers, wrong answerers, and to find stumbling points of wrong answerers, are drawn. The subjects of the experiment are TAs in the Ruby programming course. Results show that the subjects identified the conditions of the students by seeing the state transition diagram. This suggests that the visualizing answering processes of students opens a possibility of promoting instruction for teachers and TAs.

Keywords: Programming education, state transition diagram, answering process, split-paper testing, visualization

1. Introduction

1.1 Split-paper Testing

Split-paper testing is composed of fully automatically graded questions (Nakayama, Kuno, & Kakuda, 2020). It was designed to make programming proficiency tests easy to grade. Figure 1 shows the screenshot presented to the students. A test question is given as a correct program divided into lines and indicated as a set of choices represented by a choice symbol. Furthermore, it is mixed with wrong choices. Students must drag and drop the displayed choices onto the answer form to complete the program for the given problem. A log of the choices is recorded when the student answers the question and can be collected from the web system and analyzed as personal data.


Problem Title
Program without branch (A)
Created Symbol Sequence
Code Area
Problem Text
Write a method 'calc' that takes a number x and returns the sequence of x.
Choice Area

分岐のないプログラム(A)

数xを受け取り、x^2を返すメソッドcalcを書け。

最後に結果をreturnで返すこと。コードは最短でなくてよいが、冗長な (結果が使われなかったりあとで効果が取り消されたりするような) 処理 を含めないこと。

正解:

コード領域

選択肢

ア def calc(x)
イ end
ウ y = x
エ y = y + 1
オ y = y - 1
カ y = y * x
キ return y

Figure 1. Split-Paper Question that is Presented to the Students.

1.2 Fundamental Issues of Programming Education

Many students are faced with difficulties in learning programming languages. McCracken et al. (2001) found students had not learned to write practical programs at the end of an introductory programming course. Lister et al. (2004) attributed this to a lack of problem-solving skills since most students were weak in the skills needed to solve problems. Gomes and Mendes (2007) suggested the lack of understanding of algorithms was a challenge.

Additionally, it is difficult to design an optimal instruction suitable for all students (Essi, Kirsti, & Hannu-Matti, 2005). However, teachers can guide students in building their knowledge and skills through carefully designed materials and approaches.

2. Related Work

2.1 Parson’s Problems

Parson’s problems, which are similar to split-paper testing, have been widely studied. There are many variations of Parson’s problems (Petri & Ville, 2011).

Generally, problems associated with programming exams are classified into two types. One is code tracing, and the other is code writing. There have been comparative studies on code tracing, code writing, and Parson’s problem. Lister et al. (2010) also found a correlation between scores on code tracing and Parson’s problem. Thus, Parson’s problem is an excellent alternative to the traditional code tracing and code writing problems.

Helminen et al. (2012) developed an automatic feedback system and analyzed the solution of Parson’s problem using a state transition diagram. Their study analyzed a conventional Parson’s problem with an indentation operation added. The feedback system sends error messages to students about fewer lines, wrong orders, or wrong indentation. From analyzing the solutions of the students, it was found that the solution paths varied. Indentations were aligned together after fragments of the code with no indentation were aligned first. The results of the analysis are used to improve the automatic feedback system.

Salil and Amruth (2020) proposed an analysis method that focused on the distance between solution paths of Parson’s problems. The analysis was based on the edit distance, which is a modification of the Levenshtein distance. Results showed that students were inclined to difficulties in the few steps leading to the solution.

2.2 Jigsaw Code

Yamaguchi and Oba (2020) developed the Jigsaw Code, which is similar to split-paper testing. Jigsaw Code is analyzed using co-occurrence matrices and found that when two lines are selected almost simultaneously, some relationships existed among them. Using co-occurrence matrices, the jigsaw understood the strategies of students, such as the tendency to place the for-loop chunks first.

2.3 Relationship to This Study

The current study provides direct support for instructing teachers and TAs. Promoting the instruction of teachers and TAs reduces the number of students with difficulties. We analyzed and visualized the answering processes of students using the state transition diagram. The state transition diagram is shown to teachers and TAs during programming courses. This study is positioned as a basic analysis to determine how to visualize the information as the state transition diagram from the analysis of the answering process.

3. Methods

3.1 Preliminary Analysis to Prepare for the Experiment

The method used to analyze the data from the state transition diagram will be designed and discussed. We consider a set of exam questions answered by students for the analysis. The method and results of the analysis are explained in this subsection.

3.1.1 Target of Analysis

A class of 62 (50 males, 12 females, mean age 19.6 years, variance 1.07) fisrt-year university students, in a science and technology university offering a first-year course in programming is our target for analysis. Ten questions in the actual examinations for the Ruby programming language are analyzed with all questions from split-paper testing.

The history of operations, including adding, deleting and changing orders of choices are recorded as logs when students attempt split-paper testing. The CSV files for each student contain the times and answer sequences of all questions, respectively. The sentence and choice for every question as well as the correct answers can be obtained as an XML file. The CSV and the XML files are proceeded by analyzing scripts. Graphviz is used to draw a state transition diagram.

3.1.2 Methodology of analysis

For each question, different types of state transition diagrams are generated for each purpose of analysis. The present study, draws the following three types of state transition diagrams: state transition diagram of correct answerers, state transition diagram of wrong answerers, and state transition diagram to find stumbling points of wrong answerers. These state transition diagrams are created by integrating the individual answers of all students.

State transition diagrams are composed of nodes and paths with either a PDF or PNG file generated as an output of each state transition diagram. A node represents the history of each choice from the initial state (blank) to the final state (answer to be submitted). The start node and the correct answer nodes are represented in gray color. The line weight of the path is proportional to the number of transitions from one node to the next. The line weight of the pass is expressed by the thickness. The label of the path indicates the number of transitions, which is the number of students who get through the path. The size of every node in the state transition diagram of correct answerers and wrong answerers is fixed.

We introduced a stumbling node in the state transition diagram to find stumbling points of wrong answerers. The stumbling node is intended to make it easier for teachers and TAs to identify when the state transition diagram is seen. A stumbling node is defined as a node with two or more wrong answerers among the nodes with correct answerers. The size of the stumbling node is expanded by changing the node radius. It is proportional to the product of the percentage of correct answerers and the number of wrong answerers in the node. Suppose the size of the stumbling node is proportional to the number of wrong answerers alone, then the stumbling node will be oversized, and the graph will be biased. Additionally, a low percentage of correct answerers can have many wrong answerers, which results in a large number of stumbling nodes, which in turn may cause a bias in the graph. The stumbling node is represented by light gray color. Moreover, the color paths are represented in 10 levels according to the number of correct answerers.

3.1.3 Results and Discussion of analysis

Figure 2 is a state transition diagram for a particular problem. From the state transition diagrams of correct answerers, questions with a high percentage of correct answerers are answered in the same order as the correct answerers. This can be because they imagine the whole program. In addition, several correct answerers chose Def and End at first, indicating that the outline of the program was first assembled, then contents were considered. These results may be a useful point for instruction.

The state transition diagrams of wrong answerers show no trend compared with the state transition diagrams of correct answerers. However, some choices lead to the wrong answers. Some of the wrong answerers excluded important descriptions on the control structure, such as End and Return. In order to eliminate simple but important mistakes such as forgetting to insert End, it is considered a key point of instruction to assemble the outline of the program first.

From the state transition diagram to find stumbling points of wrong answerers, it was found that for questions with a low rate of correct answerers, there was little information available in this state transition diagram. In some cases, the red paths coming from the stumbling nodes were connected to another stumbling node. Several choices are important, unlike the correct answerers, the wrong answerers added unnecessary choices or removed necessary choices. The state transition diagrams to find stumbling points of wrong answerers are more difficult to grasp than those of the correct and wrong answerers, respectively. However, we found that node deviated from paths to the correct answer and what choice was wrong.

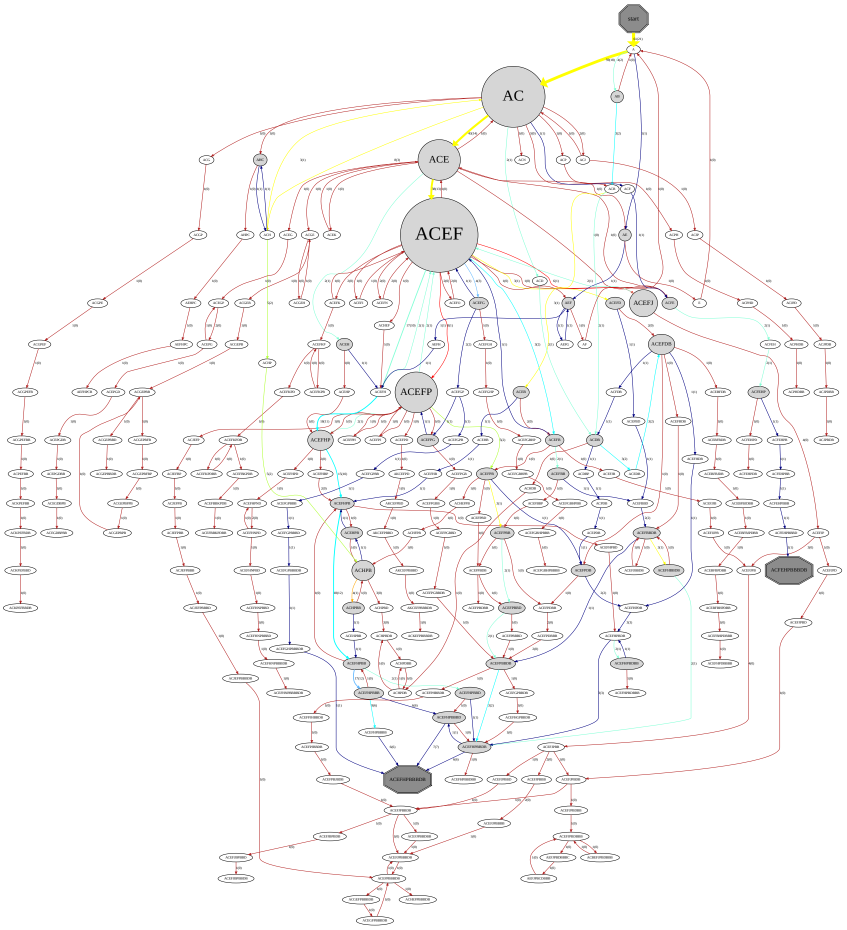


Figure 2. State Transition Diagram to Find Stumbling Points of Wrong Answerers.

3.2 Experiment Design

Figure 3 shows a scene of the experiment. The subjects of the experiment are five graduate students with experience as TAs in the programming course. The subjects examine the state transition diagrams shown on a monitor. It is allowed for the subjects to operate the answer interface using a laptop. They note on the writing paper the findings achieved by examining the state transition diagrams. The questions analyzed are four questions in Ruby language. The percentages of correct answerers for each question differed. This experiment detects the stumbling points of the wrong answerers in the state transition diagrams. It allowed subjects to search for nodes in the state transition diagrams generated as PDF files with inputting answer sequences. The answer interface is the same as the students', allowing the subjects to drag and drop the choices. However, correct answers were displayed at the start. The subjects noted the stumbling points of the wrong answerers, details, and reasons for the wrong answer. After the experiment, questionnaires and interviews were administered.

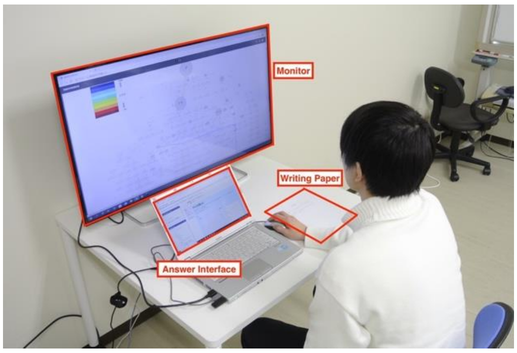


Figure 3. Scene of the Experiment.

4. Results and Discussion

The stumbling points of wrong answerers noted by each subject agree on the same points. The method to visualize the stumbling points is confirmed to be understood correctly.

The list and the results of the questionnaire are shown in Table 1 and Table 2, respectively. All the questions were asked on a Likert scale (1: completely disagree, 2: disagree, 3: undecided, 4: agree, 5: strongly agree). Q1-1 to Q1-4 is asked for each state transition diagram. The average values of the results of Q1-1 to Q1-4 are shown in the Table 2. The Cronbach's alpha coefficient for these questionnaires was 0.86.

Table 1. *The list of the Questionnaire*

Number	Question
Q1-1*	I could read the situation of the answer from the state transition diagram.
Q1-2*	The information shown in the diagram was sufficient.
Q1-3*	I could understand the points where the wrong answerers stumbled from the diagram.
Q1-4*	I was able to understand the contents of the program that caused the students to stumble.
Q2-1	I was interested in the students' solving process.
Q2-2	I saw a trend in their answerers.
Q2-3	I found the diagrams useful in teaching students.
Q2-4	I wanted to use the diagrams in my teaching.

* These questionnaires were asked for each of the four questions.

Table 2. *The Results of the Questionnaire*

Number	TA1	TA2	TA3	TA4	TA5	Ave.	S.D.
Q1-1	4.00	4.00	4.25	3.00	3.50	3.75	0.20
Q1-2	3.50	4.00	4.00	3.50	5.00	4.00	0.30
Q1-3	4.00	3.50	3.75	2.75	3.25	3.45	0.19
Q1-4	4.25	4.75	4.00	2.50	2.50	3.60	0.87
Q2-1	4.00	5.00	5.00	5.00	4.00	4.60	0.49
Q2-2	4.00	2.00	4.00	4.00	3.00	3.40	0.80
Q2-3	4.00	4.00	4.00	3.00	2.00	3.40	0.80
Q2-4	4.00	4.00	5.00	3.00	1.00	3.40	1.36

Table 2 affirms the information displayed in the diagrams as sufficient throughout the experiment. Since subjects showed interest in the answering process of students through this experiment, then visualizing the answering process can be useful for TAs. Although the tendency of the answering process is difficult to grasp, using the current state transition diagram to support the students is TA dependent.

From the interview, many subjects found the stumbling points of wrong answerers by picking up larger stumbling nodes. A few subjects said that it was difficult to learn from the current state transition diagram. Others said that they figured out the correct answerers by looking at the blue bold line in the diagram. As for the answering process of the wrong answerers, it was difficult for the subjects to grasp because of many branches. Generally, many subjects affirm that the current state transition diagram was appropriate for visualizing the answering process. However, a few subjects did not agree. The reason for their position is considered as the amount of information in the state transition diagram being too large and difficult to judge. For improvement, there have been suggestions to reduce the amount of information in the state transition diagram. The current state transition diagram, displays the number of students transitioning between nodes even if it is one student. It is considered as a way of introducing a function such as clustering or turning off nodes with many branches. Another suggestion for improvement is to display the total number of students in the stumbling nodes. Therefore, the current state transition diagram has room for further improvement to make it easier for teachers and TAs to employ state transition diagrams for instruction.

5. Conclusion

The present study visualizes the answering process of students in split-paper testing for novice programmers using a state transition diagram. Analysis and evaluation experiments were conducted. From the analysis, students who answered correctly tended to make up methods, such as Def, End, etc. If the number of correct answerers to a question was high, the answer was assembled in the correct order. For wrong answerers, such students excluded important descriptions for the control structure, such as End and Return. Thus, it is a key point of instruction to assemble the outline of the program first. From the results of evaluating the experiment, it was found that the current state transition diagram has sufficient information on the individual answering process. However, it is difficult to grasp the overall tendency and stumbling points from the information. As pointed out in the interviews, there is room for improvement in the current state transition diagram. This visualization method would be a sophisticated instructing tool for teachers and TAs.

Acknowledgements

This work has been partly supported by the Grants-in-Aid for Scientific Research No. JP19H01710 by MEXT (Ministry of Education, Culture, Sports, Science and Technology) in Japan.

References

McCracken, M., Almstrum, V., Diaz, D., Guzdial, M., Hagan, D., Kolikant, Y. B.-D., Laxer, C., Thomas, L., Utting, I., & Wilus, T. (2001). A Multi-National, Multi-Institutional Study of Assessment of Programming Skills of First-Year CS Students. ACM SIGCSE Bulletin, 33(4), 125–180.

Lister, R., Adams, E. S., Fitzgerald, S., Fone, W., Hamer, J., Lindholm, M., McCartney, R., Moström, J. E., Sanders, K., Seppälä, O., Simon, B., & Thomas, L. A Multi-National Study of Reading and Tracing Skills in Novice Programmers. ACM SIGCSE Bulletin, 36(4),119–150, 2004.

Gomes, A. & Mendes, A.J. (2007). Learning to program - difficulties and solutions. 283-287.

Lahtinen, E., Ala-Mutka, K., & Järvinen, HM. (2005). A study of the difficulties of novice programmers. ACM SIGCSE Bulletin. 37. 14-18.

Nakayama, Y., Kuno, Y., & Kakuda, H. (2020). Split-Paper Testing: A Novel Approach to Evaluate Programming Performance. Journal of Information Processing. 28. 733-743.

Ihantola, P. & Karavirta, V. (2011). Two-Dimensional Parson's Puzzles: The Concept, Tools, and First Observations. Journal of Information Technology Education: Innovations in Practice. 10. 1-14.

Lister, R., Clear, T., Bouvier, D., Carter, P., Eckerdal, A., Jackovà, J., Lopez, M., McCartney, R., Robbins, P., Seppälä, O., & Thomas, E. (2010). Naturally Occurring Data as Research Instrument: Analyzing Examination Responses to Study the Novice Programmer. ACM SIGCSE Bulletin, 41(4),156–173.

Helminen, J., Ihantola, P., Karavirta, V., & Malmi, L. (2012). How Do Students Solve Parsons Programming Problems? - An Analysis of Interaction Traces. ICER'12 - Proceedings of the 9th Annual International Conference on International Computing Education Research. 119-126.

Maharjan, S. & Kumar, A. (2020). Using Edit Distance Trails to Analyze Path Solutions of Parsons Puzzles, Educational Data Mining 2020 (EDM 2020). 638-642.

Yamaguchi, T. & Oba, M. (2020). Measurable Interactive Application to Find Out User Recognition and Strategy when Problem Solving. Journal of Software. 12-22.

Investigating Relevance of Prior Learning Data Connected through the Blockchain

Patrick OCHEJA[a*], Brendan FLANAGAN[b], Rwitajit MAJUMDAR[b] & Hiroaki OGATA[b]
[a]*Graduate School of Informatics, Kyoto University, Japan*
[b]*Academic Center for Media and Computing Studies, Kyoto University, Japan*
*ocheja.ileanwa.65s@st.kyoto-u.ac.jp

Abstract: Learners often change their learning environment over the course of their education. This makes it difficult to measure their engagement across different contexts due to a lack of seamless connection and shared analytics across heterogenous learning systems. Previous research has shown that access to prior engagement information of learners can be useful in enabling personalization, learning content design and early identification of problematic prerequisite topics. In this paper, we connect learning systems at two different schools through the blockchain to enable the transfer of learning footprints across both schools. Our primary aim is to investigate the relevance of students' prior engagement behaviour and provide stakeholders with actionable insights on dashboards. Specifically, we analyze the engagement behaviour in Junior High School grade 3 Math course of students who are currently in High School grade 1. Engagement in this context is defined based on five metrics: self-evaluation, cognitive behaviour, backtracking behaviour, time commitment and content completion rate. We further validate relevance by measuring the correlation between students' engagement level and their final score. Our analysis shows a significant difference in mean scores of very high and very low engagement students. Also, for each of the courses and scores, we provide stakeholders access to the learning materials used, assessments taken and the solutions by the students. Finally, we present implications for the field and present potential directions on how to use decentralized learner data to improve learning outcomes.

Keywords: Blockchain, education, engagement, learning behaviour, students, lifelong learning

1. Introduction

Teachers often face a common problem of not knowing the past learning engagements of their students. While final grades or scores may be contained in academic transcripts, it is difficult to measure students' engagement from transcripts. Trowler (2010) defines student engagement as the interaction between the time, effort and other relevant resources invested by students and institutes towards optimizing learning experience and to enhance students' performance. The differences in learning purposes, preferences, and motivations of students can result in different types of engagement behaviour during learning which may in-turn affect their performance (Li & Tsai, 2017). Previous research has shown that students' engagement in the learning environment is closely related to their learning outcome (Hu & Li, 2017; Lu, Huang, Huang & Yang, 2017). Thus, giving teachers access to their students past engagement could equip them with information about the possible challenges students may face, eliminate repetitive learning, how to adapt learning contents and provide support to students with prior low engagement.

To measure students' engagement at different times, it becomes necessary to access and analyze their total experience while learning at an institute. However, access to students' learning data after they change school is often difficult. This is largely attributed to the heterogeneous nature of learning systems and the lack of transferability of lifelong learning logs across schools (Baker, 2019). The advent of decentralized technologies such as the blockchain opens up new ways to address this problem. Ocheja, Flanagan, & Ogata (2018) proposed a blockchain of learning logs platform (BOLL) that can connect learning behaviour logs of students across different schools on a secure and immutable ledger. While the BOLL system solves the problem of learning data continuity, this paper presents a first of its kind research on providing teachers access to insights drawn from their students prior learning data such as engagement and learning outcome. In this work, we use the BOLL platform to provide teachers access to their students past learning engagements and investigate the relevance of students' past learning behaviour logs. For example, when students move from JHS 3 to HS 1, their HS 1 teacher is given access to the students' past learning behaviour logs. However, the teacher does not have data analytics skills to know if the learning behaviour logs have any effect on the final scores obtained. Our main argument is that it is not enough to provide access to past learning logs: the relevance of such data should also be communicated to the stakeholders. This is important because in most cases, stakeholders do not have the required data analytics to carry out such investigations on their own. We also provide a first of its kind access to the learning materials and assessment data (questions, students' and teachers' solutions) used by the student at their previous school using the marketplace (Boll-M) feature of Boll (Ocheja, Flanagan, & Ogata, 2019a). Specifically, this paper is focused on answering the following research questions:

RQ1. What are the engagement levels of students at a past learning environment?

RQ2. How relevant are these engagements to students' past learning outcome?

RQ3. How can teachers access additional information about learning outcomes?

The rest of this paper is organized as follows: In the second section, we review related works on student engagement analysis, investigate the correlation between engagement, and performance and highlight the originality of this work. The third section introduces our research methodology and the processes involved in retrieving the learning behaviour logs of students from two different schools on BOLL using the same blockchain identity, and how we calculate the engagement metrics. We present the results from our analysis and visualizations for stakeholders in the fourth section. Finally, in section five, we discuss the key findings of this research, open challenges, possible solutions and future work.

2. Related Work

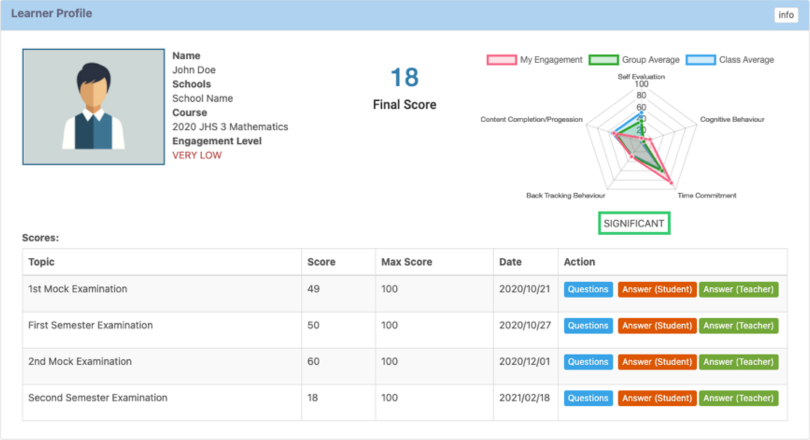
Fredricks, Blumenfeld and Paris (2004) classified student engagement in three dimensions: Behavioural, Emotional, and Cognitive. Behavioural engagement entails students' participation in learning, academic tasks and school activities, positive conduct, and absence of disruptive behaviours (Fredricks et al, 2004). Emotional engagement deals with students' affective reactions in class such as interest, boredom, happiness, sadness, and anxiety (Connell & Wellborn, 1991; Skinner & Belmont, 1993). Cognitive engagement refers to students' motivation, effort and strategic use of provided learning resources through different methods such as self-regulation and meta-cognition (Fredricks et al, 2004). This work focuses on measuring students' behavioural and cognitive engagement from past and current learning behaviour logs of students across schools and platforms.

Most of the previous studies mainly investigated students' engagement using data from the current learning environment (Li & Tsai, 2017; Lu et al., 2017; Vytasek, Patzak & Winne, 2020). While these past studies have provided useful results for the problems they addressed, we argue that students' prior engagement can provide additional important information very early: solving the *cold-start problem*. This could help learning analytics systems to make more effective personalization decisions such as recommendations, and learning preference settings. Also, because students' prior engagement and achievement are predictive of their subsequent goals (Martin & Liem, 2010), providing teachers and students with such information becomes useful especially for teaching and self-regulation (Boekaerts & Corno, 2005). Access to learner data across different institutes they have attended is still lacking and this research makes the first practical effort to facilitate transfer of learner data across schools and measure its impact on teaching and learning outcomes.

Our unique contribution in this research is to make students' prior engagement accessible when students change school. Teachers at their new school can then access and use insights from such data to improve their teaching, students' engagement and learning outcome. It is important to note that this research does not propose a new method of measuring student engagement: our main focus is to use existing techniques to measure students' engagement based on their learning behaviour logs at their previous school.

3. Visualization of Prior Learning Data

We implemented four different visualizations for stakeholders to view the past engagement of students. The *learner profile* shown in Figure 1 gives a comprehensive summary of a student's past engagement and their achievements. This can also tell the teacher if the past engagement is correlated to the student's score or not. For each of the assessments, one can also view the student's solution as well as the correct solution. The *Engagement Transition* in Figure 2 gives stakeholders ability to view change in engagement level of a group of students using iSAT (Majumdar & Iyer, 2016). For example, teachers can check transition across a period of time to know when (or at what point in the past) a student's learning behaviour changed (improved or needs intervention). The teacher can also compare engagement changes across courses, contents or activities.



Learner Profile

info

Name
John Doe
Schools
School Name
Course
2020 JHS 3 Mathematics
Engagement Level
VERY LOW

18
Final Score

Scores:

Topic	Score	Max Score	Date	Action
1st Mock Examination	49	100	2020/10/21	Questions Answer (Student) Answer (Teacher)
First Semester Examination	50	100	2020/10/27	Questions Answer (Student) Answer (Teacher)
2nd Mock Examination	60	100	2020/12/01	Questions Answer (Student) Answer (Teacher)
Second Semester Examination	18	100	2021/02/18	Questions Answer (Student) Answer (Teacher)

Figure 1. Learner Profile.

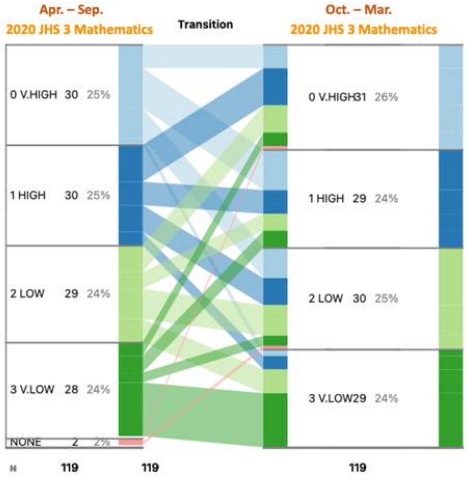


Figure 2. Temporal Change in Engagement Level.

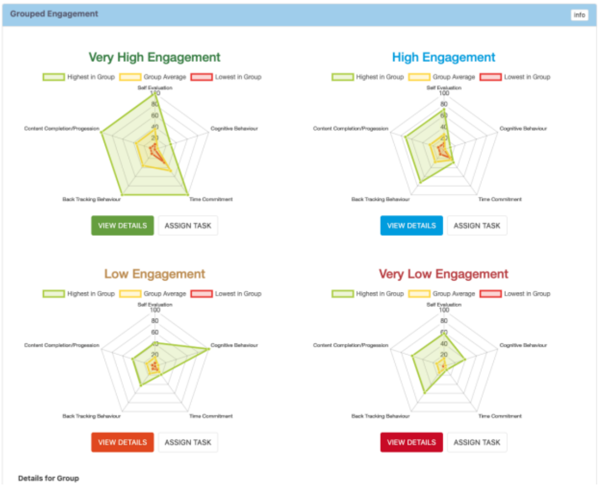
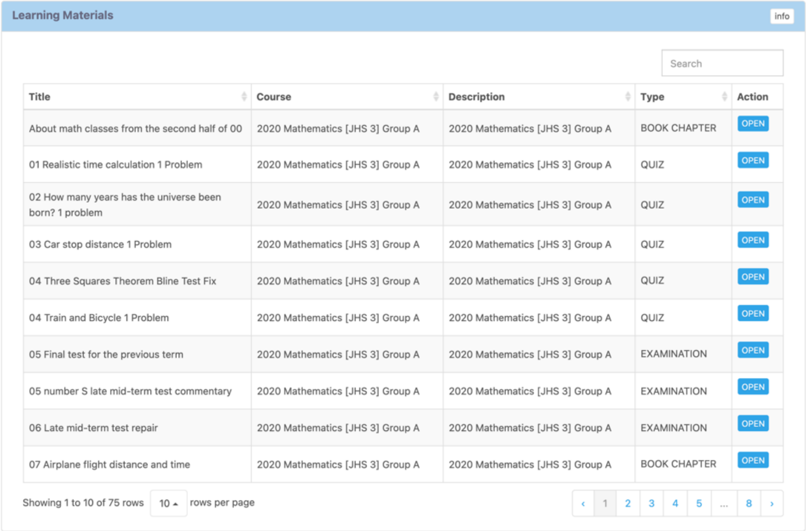


Figure 3. Detail profiles of Engagement Groups.

The *engagement groups* visualization in Figure 3 enables stakeholders to view engagement profile of different engagement cohorts in the class and to know what characteristic are prominent among different cohorts. One can also view the details of each student in each cohort and assign specific tasks such as revisions and assessment retake. The *learning materials* interface show in Figure 4 provides stakeholders a way to access the learning materials students have used in the past including: textbooks, quiz questions, students' solutions and lecture slides. Figures 1 – 4 are from a real implementation of the Boll system currently deployed at a school in Japan.

4. Research Method

In this research, we use the Boll system (Ocheja, Flanagan, Ueda & Ogata, 2019) to connect the learning behaviour logs of students across two schools in Japan. We first setup the Boll system, connect it to the Learning Records Store (LRS) of the Junior High School (JHS) and assign a blockchain address to each student. The Boll system also keeps track of each student's ID at that school. This is then used to identify the records to be transferred when the student change school. When students in these schools move from the JHS 3 to High School (HS) 1 (a different school), we also transfer their past learning logs on the BookRoll system (Flanagan & Ogata, 2018) to their new school. The HS also has a similar setup of the Boll system with connections to the LRS.



Title	Course	Description	Type	Action
About math classes from the second half of 00	2020 Mathematics [JHS 3] Group A	2020 Mathematics [JHS 3] Group A	BOOK CHAPTER	OPEN
01 Realistic time calculation 1 Problem	2020 Mathematics [JHS 3] Group A	2020 Mathematics [JHS 3] Group A	QUIZ	OPEN
02 How many years has the universe been born? 1 problem	2020 Mathematics [JHS 3] Group A	2020 Mathematics [JHS 3] Group A	QUIZ	OPEN
03 Car stop distance 1 Problem	2020 Mathematics [JHS 3] Group A	2020 Mathematics [JHS 3] Group A	QUIZ	OPEN
04 Three Squares Theorem Bline Test Fix	2020 Mathematics [JHS 3] Group A	2020 Mathematics [JHS 3] Group A	QUIZ	OPEN
04 Train and Bicycle 1 Problem	2020 Mathematics [JHS 3] Group A	2020 Mathematics [JHS 3] Group A	QUIZ	OPEN
05 Final test for the previous term	2020 Mathematics [JHS 3] Group A	2020 Mathematics [JHS 3] Group A	EXAMINATION	OPEN
05 number S late mid-term test commentary	2020 Mathematics [JHS 3] Group A	2020 Mathematics [JHS 3] Group A	EXAMINATION	OPEN
06 Late mid-term test repair	2020 Mathematics [JHS 3] Group A	2020 Mathematics [JHS 3] Group A	EXAMINATION	OPEN
07 Airplane flight distance and time	2020 Mathematics [JHS 3] Group A	2020 Mathematics [JHS 3] Group A	BOOK CHAPTER	OPEN

Showing 1 to 10 of 75 rows 10 rows per page

Figure 4. Past Learning Materials Transfer across Schools.

For this study, we analyzed the learning behaviour logs of 109 students in JHS 3 Mathematics course in 2020 academic year who are currently in HS 1 and have enrolled in the HS 1 Mathematics course in 2021 academic year. Our analysis includes: engagement behaviour cohorts, temporal and spatial change in engagement and learning contents visualization. We measure engagement as a sum of different student behaviours categorized in to 5 dimensions: self-evaluation (S_e), cognitive behaviour (C_b), backtracking behaviour (B_b), time commitment (T_c) and content progress/completion (C_p). We define self-evaluation (S_e) as the students' ability to evaluate correctly their own solution to quiz questions. S_e is calculated as a fraction of the quiz answers from the student which were correct and rightly marked as correct by the student. Cognitive behaviour (C_b) is a measure of the students' cognitive action through cognitive indicators such as yellow and red markers added on learning materials through the BookRoll system (Akçapinar, Hasnine, Majumdar, Flanagan & Ogata, 2019). The backtracking behaviour (B_b) is an indication of how often students revisit concepts in order to improve their understanding or master such concepts. This is calculated as a weighted sum of total previous page visit actions divided by the total next page visit actions and the total previous page visit actions (Yang, Chen, & Ogata, 2021). Time commitment (T_c) is a measure of how often students study and it is calculated as the weighted sum of the total time, total number of content usage events and the total number of unique days students used the contents of the course. Content progress/completion (C_p) is a measure of how students advance towards completing the study materials. It is calculated as the weighted sum of total open and next page actions and total sum of long and short events. It is important to note that the parameters of each engagement metric were percentile rank of their actual values. Thus, student overall engagement is calculated as:

$$\text{Engagement} = S_e + C_b + B_b + T_c + C_p$$

In table 1, we show a summary description of the dataset for the JHS 3 Math course in 2020. The engagement metrics previously discussed were extracted from the dataset of the students who took the final exam and were graded. The engagement score was used to divide into quartile groups of 4 different engagement levels: Very High (≥ 75th percentile), High (≥50th percentile), Low (≥25th percentile) and Very Low (< 25th percentile) using percentile rank. We then proceeded with ensuring the data meet the assumptions of a one-way Analysis of Covariance (ANOVA) before performing conducting a test for a significant difference in the mean score for each engagement level. Finally, we developed 4 visualizations for teachers to view students' past engagement showing information such as: learner profile, group engagement, temporal and spatial engagement change and learning materials used.

Table 1. *Description of the Dataset*

	No. of Students	Total Logs	No. of Students graded
Group A	40	123,678	38
Group B	40	98,080	38
Group C	40	125,619	33

5. Results

Before carrying out an Analysis of Covariance (ANOVA) between the engagement levels and score, a Shapiro-Wilk test was conducted to determine the normality of the data. The result (0.99, $p > 0.05$) revealed that the score data across the different engagement levels followed a normal distribution. A further test for homogeneous variance using Levene's test indicated homogeneity of variances across the different engagement levels (F (3,105) = 2.272, $p > 0.05$). We then conducted a parametric one-way ANOVA to determine whether the mean scores of all engagement levels are different. The result (F(3,105) = 3.783, $p < 0.05$) indicated a significant difference in the mean scores for all engagement levels. A further post-hoc test using the Games-Howell test (due to unequal sample sizes) showed that the difference between very high and very low engagement levels is significant ($p < 0.05$) as presented in table 2. The implication of this result is that very low and very high engagement levels are indicative of the final performance of students and provide actionable insights for guiding future teaching and learning.

Table 2. *Post-Hoc Test (Games-Howell) Results of Scores between Engagement Levels (Mean Difference, Standard Error)*

	N	Score (μ)	SD	Very High	High	Low	Very Low
Very High	28	59.64	16.01	-	3.72 (4.12)	6.50 (4.02)	14.53 (5.06)*
High	27	55.93	14.51		-	2.78 (3.85)	10.82 (4.92)
Low	27	53.15	13.76			-	8.04 (4.84)
Very Low	27	45.11	21.07				-

*Note. * $p < 0.05$.*

6. Discussions

This work makes an important contribution of investigating and informing stakeholders the effect of students' prior engagement on their final scores at a different learning environment. Such information makes it possible for teachers to provide specific interventions at the start of a new class without having to wait to collect some data in the first few weeks. Although the results from our analysis only revealed a significant correlation between the scores and engagement of very high and very low engagement students, we propose this type of analysis to be performed when providing stakeholders with learning logs from a different learning environment.

In addition to engagement and final scores, this work provided access to resources such as the students' solution to examination questions and learning materials used. Access to this type of data give teachers additional information about the students' ability, and challenges with respect to the assessment questions. We acknowledge that in some cases, other contextual information may be required to correctly interpret the engagement measures extracted from the learning logs. Also, students may have received other scores different from the final score. It may be useful to consider how the students' engagement at intervals preceding other assessment affected their performance.

7. Conclusion

The transfer of learning logs and materials across different schools is an important requirement to solving the cold-start problem. This work presented metrics for defining engagement levels of students transitioning different learning environments. To validate the usefulness of the transferred data, we conducted some statistical tests which showed that the proposed engagement measures had a significant impact on students' final score. We also presented how teachers can access additional data from the students past learning data such as quiz answers by students, and the textbooks used as well as the previous teachers' lecture materials. Future work will be focused on validating the usefulness of the proposed visualizations with key stakeholders and the impact of our system.

Acknowledgements

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B)20H01722, JSPS Grant-in-Aid for Scientific Research (S)16H06304 and NEDO Special Innovation Program on AI and Big Data 18102059-0.

References

Akçapinar, G., Hasnine, M. N. , Rwitajit, M., Flanagan, B., & Ogata, H., (2019). Exploring the Relationships between Students' Engagement and Academic Performance in the Digital Textbook System. 27th *International Conference on Computers in Education*, Taiwan.

Baker, R. S. (2019). Challenges for the future of educational data mining: The Baker learning analytics prizes. *Journal of Educational Data Mining*, *11*(1), 1-17.

Boekaerts, M., & Corno, L. (2005). Self- regulation in the classroom: A perspective on assessment and intervention. Applied Psychology, 54(2), 199-231.

Flanagan, B., & Ogata, H. (2018). Learning analytics platform in higher education in Japan. *Knowledge Management & E-Learning: An International Journal*, *10*(4), 469-484.

Hu, M., & Li, H. (2017). Student engagement in online learning: A review. In 2017 *International Symposium on Educational Technology*, 39-43. IEEE.

Li, L. Y., & Tsai, C. C. (2017). Accessing online learning material: Quantitative behavior patterns and their effects on motivation and learning performance. *Computers & Education*, 114, 286-297.

Lu, O. H., Huang, J. C., Huang, A. Y., & Yang, S. J. (2017). Applying learning analytics for improving students engagement and learning outcomes in an MOOCs enabled collaborative programming course. *Interactive Learning Environments*, 25(2), 220-234.

Majumdar, R., & Iyer, S. (2016). iSAT: a visual learning analytics tool for instructors. *Research and Practice in Technology Enhanced Learning*, 11(1), 1-22.

Martin, A. J., & Liem, G. A. D. (2010). Academic personal bests (PBs), engagement, and achievement: A cross-lagged panel analysis. *Learning and Individual Differences*, 20(3), 265-270.

Ocheja, P., Flanagan, B., & Ogata, H. (2018). Connecting decentralized learning records: a blockchain based learning analytics platform. In Proceedings of the 8th *international conference on Learning Analytics and Knowledge*, 265-269.

Ocheja, P., Flanagan, B., & Ogata, H. (2019a). Decentralized E-Learning Marketplace: Managing Authorship and Tracking Access to Learning Materials Using Blockchain. In *International Cognitive Cities Conference*, 526-535. Springer, Singapore.

Ocheja, P., Flanagan, B., Ueda, H., & Ogata, H. (2019). Managing lifelong learning records through blockchain. *Research and Practice in Technology Enhanced Learning*, 14(1), 1-19.

Vytasek, J. M., Patzak, A., & Winne, P. H. (2020). Analytics for student engagement. In *Machine learning paradigms*, 23-48. Springer, Cham.

Yang, C. C., Chen, I. Y., & Ogata, H. (2021). Toward Precision Education: Educational Data Mining and Learning Analytics for Identifying Students' Learning Patterns with Ebook Systems. *Educational Technology & Society*, 24(1), 152-163.

Analysing Reachable and Unreachable Codes in App Inventor Programs for Supporting the Assessment of Computational Thinking Concepts

Siu Cheung KONG[a*], Chun Wing POON[b] & Bowen LIU[b]
[a]*Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong*
[b]*Centre for Learning, Teaching and Technology, The Education University of Hong Kong, Hong Kong*
*sckong@eduhk.hk

Abstract: Examining students' programming artefacts is a practical and effective approach to assessing the development of students' computational thinking (CT). Some tools can automatically assess students' programming artefacts according to predefined rubrics. However, these tools may fail to properly assess students' artefacts when unreachable codes exist. In this study, we designed an automatic assessment tool that can analyse reachable and unreachable codes independently in students' App Inventor programs in static code analysis. Our tool counts reachable and unreachable codes separately to avoid potential bias. The designed tool can process students' programs in a batch and calculate the average block frequency of all programs in the batch. The average block frequency provides teachers with an initial impression of the achievement of the entire class. Previous assessment tools have used predefined rubrics to assess students' performance. We suggest that students' understanding of CT concepts could be assessed by block frequency comparison with a known programming solution. By comparing the frequency of the reachable and unreachable code in the programs with that of the starter code, students' progression in the task can be evaluated. By comparing the block frequency of these programs with a suggested solution, students' difficulties in programming can be identified.

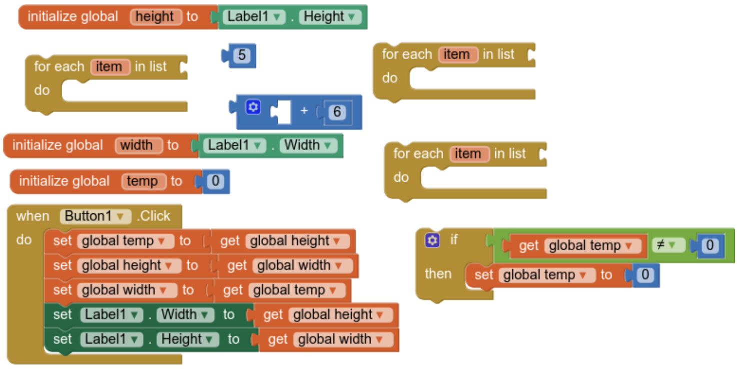
Keywords: App Inventor, assessment, computational thinking, concepts, reachable codes, unreachable codes

1. Introduction

Assessment is an important issue in computational thinking (CT) education for K–12 students. Effective and efficient CT assessments can measure students' learning and highlight gaps in their understanding. In this study, we used an enhanced static code analyser with unreachable code detection to assess students' understanding of the CT concepts in well-structured, open-ended programming tasks. We developed a web-based tool to analyse students' App Inventor programs. This tool generates a block frequency report in which reachable and unreachable code blocks are counted separately. Different from previous assessment tools that have relied on predefined rubrics for assessment, it is possible to assess students' CT concepts by directly comparing block frequency in a well-structured, open-ended programming task, which usually provides a starter code and a suggested solution. In this study, we provide an example of the application of our assessment tool to a group of students' App Inventor programs to examine students' understanding of the list data structure. With the batch processing function of this tool, teachers can easily obtain an impression of the learning status of the entire class and identify students' learning gaps in the acquisition of certain CT concepts.

2. Research Background

Computational thinking, which is a set of analytical thinking skills that uses fundamental computing concepts and practices for problem-solving, has been regarded as a necessary ability in the twenty-first century (Wing, 2006). Therefore, much effort has been devoted to promoting CT education in K–12 education (Grover & Pea, 2013; Lye & Koh, 2014). Because the goal of CT goes beyond teaching students to be digital consumers and aims to encourage problem solving with computing methods, CT education usually involves activities that provide students opportunities to create computational artefacts. Teaching CT through a programming activity is a popular and practical strategy. Students can be exposed to CT concepts during the process of creating their programmes within block-based visual programming environments. Students' programming artefacts are natural and informative products for assessing CT development. Several automatic tools, such as Dr. Scratch (Moreno-León & Robles, 2015) and CodeMasters artefacts (Von Wangenheim et al., 2018), have been developed to assess students' artefacts. These tools conduct static code analyses of students' programming artefacts and score the programs based on predefined rubrics. The rubrics set criteria to map scores to students' CT-related abilities using features and traits of code. Although these tools can significantly alleviate teachers' workload during assessment, they are inadequate when unreachable codes exist in students' programs.



```
initialize global height to Label1 . Height

for each item in list
do

5

+ 6

for each item in list
do

initialize global width to Label1 . Width

initialize global temp to 0

for each item in list
do

when Button1 .Click
do  set global temp to get global height
    set global height to get global width
    set global width to get global temp
    set Label1 . Width to get global height
    set Label1 . Height to get global width

if get global temp ≠ 0
then set global temp to 0
```

Figure 1. A Students' App Inventor Program that Contains Unreachable Codes.

Unreachable codes, also called dead codes, are codes that are never executed in the complete run-time of a program (Amanullah & Bell, 2020). Although unreachable codes seem to have no harmful effect on programs' functionality, they increase the complexity of the programs, reduce the readability of codes, and introduce potential risks in the iterative development process of programs. In using traditional static-code analysis for CT assessment, unreachable codes can distort the assessment results and provide inaccurate feedback to teachers and students. Figure 1 depicts a student's App Inventor program that attempts to switch the width and height of a label. Notably, several loop blocks are dangling without connecting to an event handler. However, this program may mislead the assessment tool to conclude that the student has properly understood the loop concept. Although having an unreachable code seems to be a bad programming habit, some novice programmers can benefit from it (Amanullah & Bell, 2018). The unreachable codes sometimes work as a temporary workspace to store unused blocks and provide a start point for novices. The proportion of unreachable codes in students' final programming products can also indicate their achievement in learning tasks. In some well-structured, open-ended programming tasks, starter codes are provided at the beginning and are not fully executable. The unreachable codes that remain in the final products may indicate students' barriers in the programming process.

3. A Web Tool for Automatic Analysis of App Inventor Programs

To identify unreachable codes, we developed a tool for the automatic detection of unreachable codes in students' App Inventor programs. The tool is available on the web for teachers to use, and it supports batch processing of students' App Inventor programs.

.

3.1 Unreachable Codes in App Inventor Programs

App Inventor is a block-based programming environment for creating mobile applications. In most situations, codes are designed in event-driven patterns in an app, and the reachable codes are components of either an event handler, a procedure, or a global variable definition. If a block of codes has a top-level block type that is not one of the above three types, it is regarded as an unreachable code block because there are no ways to activate its execution. Figure 2a shows an example of a dangling code block that is unreachable in an App Inventor program. Additionally, empty event handlers, procedures without connected blocks, and incomplete global variable definitions (Figure 2b) were also regarded as unreachable codes in this study.

There are other scenarios that make the codes unreachable, such as an event that can never be triggered due to a missing event-driven item in the design interface in App Inventor. We did not define these codes as unreachable in this study because they can be reached and rectified after a logical structured walk-through of the program design (Lemos, 1979).

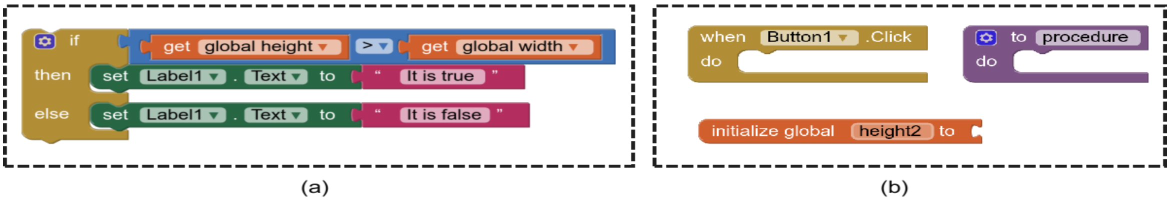


Figure 2. Examples of Unreachable Codes with 2(a) shows a dangling code block that is unreachable and 2(b) with empty event handlers, procedures without connected blocks and incomplete global variable definition

3.2 Unreachable Code Detection via XML Parsing

Our tool is an App Inventor analyser developed in Java and deployed as a web service. The web tool can analyse App Inventor programs in a batch and summarize the block frequency of all the programs. Our tool first extract the XML tree from the .aia file, then identify the blocks in the program via a lexical analysis. After identifying the blocks, our tool counts the number of blocks by traversing the XML tree and generates a block frequency report. Table 2 shows the results of analysis of a students' App Inventor program in Figure 1. The categories used in the table classify blocks according to their related CT concepts. The following two steps are used to identify whether a code block is reachable or not: 1) check whether the top-level block is either an event handler, a procedure, or a global variable definition; 2) check whether the top-level block is connected to any other blocks. Unreachable codes are not repeatedly counted in other categories.

Table 2. *The Block Frequencies of the Codes in Figure 1*

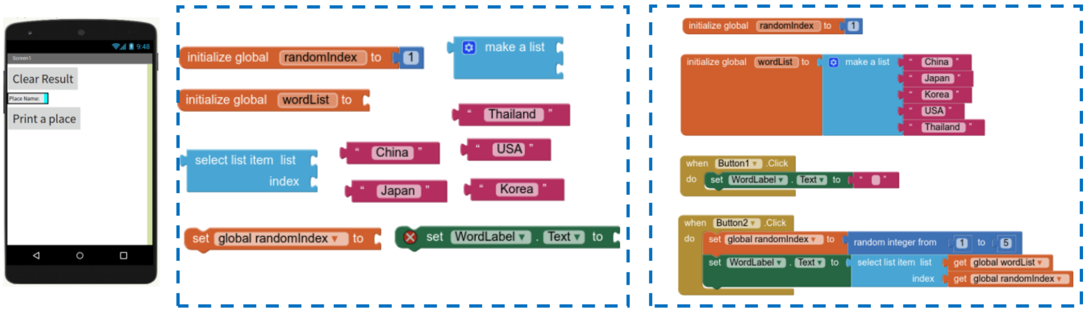
	Block Frequency
CT concepts:	
Events & Sequences	2
Conditionals	0
Operators	0
Repetition	0
Naming & Variables	14
Data Structures	0
Procedure	0
Unreachable Codes:	
Empty Event Handler	0
Empty Procedure	0
Incomplete Variable Initialization	0
Dangling Blocks	12

4. Example Analysis of a Batch of App Inventor Programs

To show the effectiveness of the assessment tool, we analysed a batch of students' App Inventor programs to illustrate how the detection of unreachable codes could indicate students' understanding of CT concepts and how teachers could assess the programs.

4.1 Data Description

Our testing data contained 27 App Inventor programs developed by primary school students. These programs were the students' assignments in an App Inventor course. In this course, the students learned how to use the list, which is an abstract data type used in programming tasks. The assignment required the students to develop a mobile app that could randomly print the name of a country from a given list. This task required students to have basic knowledge of the list data structure, including how to create a list and how to access the elements using the index of the list. The task was presented to students together with the starter code shown in Figure 3a. The starter code contained unreachable codes and acted as a start point for students to complete the programming task. A suggested solution is shown in Figure 3b.


Screen1
9:48
Clear Result
Place Name:
Print a place
initialize global randomIndex to 1
make a list
initialize global wordList to
Thailand
select list item list
index
China
USA
Japan
Korea
set global randomIndex to
set WordLabel . Text to
initialize global randomIndex to 1
initialize global wordList to make a list
China
Japan
Korea
USA
Thailand
when Button1 .Click
do set WordLabel . Text to
when Button2 .Click
do set global randomIndex to random integer from 1 to 5
set WordLabel . Text to select list item list get global wordList
index get global randomIndex

(a) (b)

Figure 3. A Programming Task using List Data Structure. 3a: the starter code of the task; 3b: a suggested solution of the programming task.

4.2 Analysis of the Students' Programs

The batch processing function of our tool allowed us to analyse all of the programs at the same time. We uploaded all 27 programs to the analyser, and part of the generated result is shown in Table 3. The generated result contained the block frequency of the 27 programs and an average block frequency of all of the programs. The block frequencies of only five programs are presented in Table 3 for further discussion. We appended the block frequency of the starter code and the suggested solution for comparison.

From the block frequency of the starter code (Table 3), we can easily notice that most of the codes are counted into the categories under unreachable codes. A decrease in unreachable codes could be used to indicate the students' progress in this task. Eleven programs (40.7%) were free of unreachable codes. The average frequencies in the unreachable code categories were significantly lower than those measured initially, which suggests that the students correctly understood the goal of the task and demonstrated a certain degree of understanding of the list data structure. By comparing the average block frequency and suggested solution, we noticed that the average frequency of blocks in the '*Events and Sequences*', '*Naming & Sequences*', and '*Data Structures*' were lower than that in the suggested solution. We calculated the standard deviation (SD) and coefficient of variation (CV) in each category and the categories of the unreachable codes (Table 4). The '*Data Structures*' category had the largest variation among the three CT concept categories, suggesting that students had a different understanding of the data structure, which was the list structure in our study. The SD and CV of the categories under unreachable codes were relatively higher than those of the CT concepts categories. This suggests that unreachable codes can be a useful indicator of students' achievement in this task and more information can be explored in these unreachable codes.

We investigated the functionality of each program by comparing it with the block frequency of the suggested solution. If a program's block frequency was identical to that of the suggested solution, the student's program likely fulfilled the required function, although the connection between the blocks would still require a manual inspection. Such an inference would be reasonable if an open-ended question were well structured. From our data, eight programs (29.6%) were found to be identical to our suggested solution. The difference in block frequency can indicate the barrier in students' problem-solving process. We compared the block frequency of program 4 to our suggested solution and found that it had zero entries in the '*Data Structure*' category and non-zero entries in the '*Incomplete Variable Initialization*' and '*Dangling Blocks*' categories. Through a further study of the source code (Figure 4), we noticed that the student included the list-related operation blocks into the app but failed to connect those blocks correctly. This means that the student was not fully familiar with the usage of the list.

Table 3. *The Analysed Results Obtained using Our Developed Tool*

	Average Block Frequency	Starter code	App Inventor Program 1	App Inventor Program 2	App Inventor Program 3	App Inventor Program 4	App Inventor Program 5	Suggested Solution
CT Concepts:								
Events & Sequences	2.44	0	3	3	3	2	2	3
Conditionals	0	0	0	0	0	0	0	0
Operators	0	0	0	0	0	0	0	0
Repetition	0	0	0	0	0	0	0	0
Naming & Variables	11.85	2	15	14	14	5	13	14
Data Structures	1.74	0	3	2	2	0	0	2
Procedure	0	0	0	0	0	0	0	0
Unreachable Codes:								
Empty Event Handler	0.14	0	0	1	0	0	3	0
Empty Procedure	0	0	0	0	0	0	0	0
Incomplete Variable Initialization	0.22	1	2	0	0	1	0	0
Dangling Blocks	2.6	9	6	1	0	12	3	0

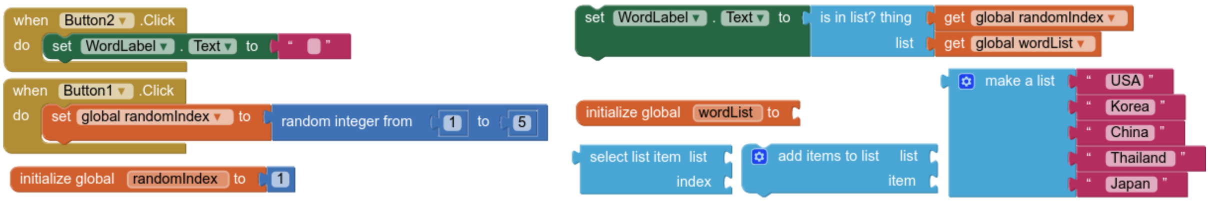

when Button2 .Click
do set WordLabel . Text to " "
when Button1 .Click
do set global randomIndex to random integer from 1 to 5
initialize global randomIndex to 1
set WordLabel . Text to is in list? thing get global randomIndex
list get global wordList
initialize global wordList to
make a list "USA" "Korea" "China" "Thailand" "Japan"
select list item list
index
add items to list list
item

Figure 4. The Block Code of the App Inventor Program 4 in Table 3.

Table 4. *The Standard Deviations and Coefficients of Variation of the Block Frequencies*

	Average Block Frequency	Standard Deviation	Coefficient of Variation

CT Concepts:			
Events & Sequences	2.44	0.974	0.398
Conditionals	0	0	-
Operators	0	0	-
Repetition	0	0	-
Naming & Variables	11.85	3.697	0.311
Data Structures	1.74	0.712	0.409
Procedure	0	0	-
Unreachable codes:			
Empty Event Handler	0.14	0.60	4.06
Empty Procedure	0	0	-
Incomplete Variable Initialization	0.22	0.50	2.27
Dangling Blocks	2.67	3.63	1.36

5. Summary

In this study, we assessed students' CT concepts by analysing reachable and unreachable code separately in a static code analysis. We defined several patterns of unreachable codes and developed an automatic analyser that could detect these unreachable codes, count the block frequency, and process students' App Inventor programs in a batch. We tested our tool on a group of students' App Inventor programs developed in a well-structured, open-ended task that required knowledge of the list data structure. By comparing the block frequency of the starter code and the average block frequency of students' programs, teachers can quickly gain a high-level impression of the learning status of an entire class. The differences in unreachable code frequencies between the starter code and students' programs indicate the students' achievement. By comparing the individual block frequency with that of the suggested solution, teachers can rapidly locate gaps in students' programming processes. The assessment tool and method used in this study require the programming tasks to be open-ended and well-structured with at least one correct solution known in advance. Our method assesses students' programs by comparing block frequencies instead of using predefined rubrics. It enables efficient assessment and customised analysis. Currently the comparison of block frequencies is performed using numeric observations. In future studies, we will consider visualising the comparison of block frequencies to enable teachers to obtain visual hints for the assessment. We will conduct experiments in primary schools to further verify the efficacy and reliability of the system. In addition to assessment, we will explore the potential of applying the system to provide self-serve real-time feedback in students' programming learning.

References

Amanullah, K., & Bell, T. (2018). Analysing students' scratch programs and addressing issues using elementary patterns. *2018 IEEE Frontiers in Education Conference (FIE)*, 1–5.

Amanullah, K., & Bell, T. (2020). Revisiting code smells in block-based languages. *Proceedings of the 15th Workshop on Primary and Secondary Computing Education*, 1–2.

Grover, S., & Pea, R. (2013). Computational thinking in K-12: A review of the state of the field. *Educational Researcher*, *42*(1), 38–43.

Lye, S. Y., & Koh, J. H. L. (2014). Review on teaching and learning of computational thinking through programming: What is next for K-12? *Computers in Human Behavior*, *41*, 51–61.

Lemos, R. S. (1979). An implementation of structured walk-throughs in teaching COBOL programming. *Communications of the ACM*, 22(6), 335-340.

Moreno-León, J., & Robles, G. (2015). Dr. Scratch: A web tool to automatically evaluate Scratch projects. *Proceedings of the 10th Workshop in Primary and Secondary Computing Education*, 132–133.

Von Wangenheim, C. G., Hauck, J. C. R., Demetrio, M. F., Pelle, R., da Cruz Alves, N., Barbosa, H., & Azevedo, L. F. (2018). CodeMaster - Automatic assessment and grading of app inventor and snap! Programs. *Informatics in Education*, *17*(1), 117–150. https://doi.org/10.15388/infedu.2018.08

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33–35.

Classification of Learning Patterns and Outliers Using Moodle Course Material Clickstreams and Quiz Scores

Konomu DOBASHI[a*], Curtis P HO[b], Catherine P FULFORD[b], Meng-Fen Grace LIN[b] & Christina HIGA[c]

[a]*Faculty of Modern Chinese Studies, Aichi University, Japan*

[b]*Learning Design and Technology, College of Education, University of Hawai'i at Mānoa, USA*

[c]*Social Science Research Institute, University of Hawai'i at Mānoa, USA*

*dobashi@vega.aichi-u.ac.jp

Abstract: In this study, learning patterns and outliers were classified using learning logs in Moodle, and a method was proposed to identify learners who were struggling in class based on the relationship between learning patterns and outliers. The proposed method utilizes the deviation between the learner's course material clickstream and the quiz score accumulated in Moodle to classify the learner into one of four learning patterns. As the number of lessons increased, many learners transitioned through four learning patterns. However, some of the top or bottom learners on the final quiz score may repeat the same learning pattern, which tends to result in outliers.

Keywords: Learning pattern, educational data mining, learning analytics, clickstream, quiz score, outlier, unsuccessful learner, Moodle

1. Introduction

Even in a class in which a large number of learners participate, more effective intervention will be possible if appropriate time-based observation of learning behavior can be performed for each individual. The main purpose of this research was to analyze the learning behavior of students and to clarify the relationship between the classification of learning patterns and the occurrence of outliers. Outliers here are quiz scores and material clickstreams that are far from average.

Learners can have various styles of learning patterns, but in this study, they were considered as follows. (1) If a learner browsed the course materials appropriately according to the progress of the lesson, the learner was considered to be highly interested in the lesson and that the lesson was taken successfully. (2) On the other hand, when the number of material clickstreams, that is, the number of material openings, was small or when the materials were not opened, the learner was considered to have low interest in the lesson and a learning pattern with insufficient engagement. (3) Furthermore, the material browsing pattern is expected to affect learners' scores on weekly quizzes and final exams.

Therefore, determining whether a learner understands the content of the material based on their scores in the quizzes and exams is possible. In this study, the learning materials' clickstream and the results of 13 quizzes were analyzed, and a method for classifying the students' learning patterns and outliers was proposed.

2. Related Research

Engagement is also measured from various aspects in motivational research in the field of education. Questionnaires are used to analyze answers to the psychological state of motivation to engage in learning activities, and question items are used as variables for measurement from multiple perspectives (Skinner et al., 2009b). Behavioral engagement has also been shown to directly regulate academic performance in individual learning situations, suggesting its importance in learning (Steinmayr et al., 2018). In this study, the learning log automatically recorded by Moodle was used to analyze learners' engagement. Behavioral engagement is defined as learning and academic engagement, including active behavior, and measures such as attention to learning and indication of effort. Therefore, clickstreams for browsing materials and quiz scores can directly define grades and are considered factors related to conventional behavioral engagement.

Many studies on learning patterns have been conducted in the past. In one case, learners completed a questionnaire, and the characteristics of their learning patterns were analyzed using factor analysis (Vermunt & Donche, 2017). Recently, research on new learning patterns utilizing a learning management system (LMS) or e-book systems has been conducted (Hsiao et al, 2019). Learning pattern analysis using an LMS or e-books involves collecting and analyzing data on learning behaviors, such as page back-and-forth movements, highlighting, underlining, and commenting (Mouri et al., 2019). Studies to detect outliers are also being conducted in the field of education. In a study of a massive open online course (MOOC), Gitinabard et al. (2018) predicted who would drop out based on student access to materials and forum logs. They were able to quickly identify individuals who were at risk of becoming unsuccessful learners, and they showed that their approach was useful for early learner intervention and guidance (Gitinabard et al., 2018).

Research is likewise being conducted on the efficacy of student support systems that integrate LMS data with student management and grading management systems. Course Signals, developed at Purdue University, is an early-intervention system that provides real-time student feedback based partly on student records accumulated in Blackboard and past learning logs (Arnold & Pistilli, 2012). A system called E2Coach at the University of Michigan sends messages to learners based on their course score data. These messages motivate learners to take the actions necessary for success, reminding them, for example, to ensure that they have sufficient time to prepare for their next exam (McKay et al., 2012). In recent years, research on data mining and dashboards that analyzes LMS learning logs has been active (Slater et al., 2017). For example, Estacio et al. (2017) used a vector space model in Moodle's learning logs to analyze the relationship between learning behavior and the final grade. They developed a method to monitor learners' behavioral levels and showed that they could find learners who are struggling in class (Estacio et al., 2017). All these studies aim to obtain useful knowledge for class management and the improvement of teaching materials.

3. Course Outline and Teaching Material

This study was tested with a class taught at Aichi University in Japan. Moodle learning logs were collected from the course "Introduction to Social Data Analysis," of which the first author was in charge at the university. Learning logs were collected from September 18, 2019 to January 8, 2020. Learners from freshman to senior can enroll in the course. There were 55.0% and 45.0% male and female undergraduates, respectively, and the age range for most of the learners was 18 to 22 years; a total of 47 learners participated in this class. The content of the class was an introduction to statistics using Excel. In the actual class, learning was undertaken over 15 weeks, starting with learning the basic use of Excel, including representative value, variance, standard deviation, simulation, frequency distribution and pivot tables, attribute correlation, covariance, correlation analysis, and regression analysis. The course materials for reading were mainly created in PDF files, comprising 12 chapters, 112 sections, 10 external URLs, and the entire material of 154 pages. The materials were divided into 112 files, and they were then uploaded to Moodle, which was set to "topic" mode.

All lessons were conducted according to the teacher's instructions; learners accessed the materials according to the instructions and learned about data processing using Excel. In the first half of the class, while reading the materials on Moodle, the learners have their Excel screens open at the same time, and they operate their personal computers to study. In the second half of the lesson, the learners carried out each exercise and were asked to submit a work file of all the exercises in Excel that they had completed. In the classroom, one material presentation monitor was prepared for every two learners, and the learners could see the materials and a demonstration of computer operations. They can also open materials from the Moodle screen on their computers at home or on their classroom computers and freely browse and download them. In this study, the number of downloads is therefore included in the material clickstream. In the first lesson of the semester, the first author explained to all learners that Moodle collects student learning logs 24 hours a day.

The clickstream in this study refers only to the number of times the course material on Moodle is opened, not to all the clickstreams for operating a personal computer. The clickstream of materials by chapter, corresponding to the question range of the quiz, is also tabulated. The clickstream of materials covered the period from the first lesson of the semester to the end of the quiz in each chapter. This log included in-class and out-of-class activity. The quizzes were conducted a week from the learning start time of each chapter. Aggregating the individual learners' material clickstreams for each learning period corresponding to the quiz was necessary, so we adopted the framework of time-series cross-section (TSCS) analysis. Excel Pivot Table can be used to aggregate the frequency distributions of multiple discrete data to generate a 2D cross-table. As the Moodle learning log contains time-stamp data, creating the TSCS table and aggregating clickstreams are possible.

Quizzes were created for each chapter using the materials that were on Moodle. We decided to adopt fill-in-the-blank-type quiz questions (i.e., complete the sentence), with five alternative answers for each question. An average of 12 quiz questions were created for each chapter, totaling 146 questions. The classes covered in this study were conducted once a week. To confirm the degree of the learners' comprehension of the lesson content in a given week, a quiz was administered the week after completing the lesson. At the beginning of the lesson, we used the quizzes on Moodle and gave 12 quizzes from weeks 3 to 15. For the weekly quizzes, learners were given five minutes to answer five questions. These questions were randomly chosen from the questions for the chapter learned in the previous week. During the final lesson of the semester, a 30-question final quiz was given. The final quiz integrated the 12 weeks of quizzes, with 30 of the 146 questions created for the weekly quizzes being randomly selected, to assess the learners' level of comprehension of the various lessons.

4. Experimental Results

4.1 Classification of Learning Patterns

This section describes how to anticipate any relationships between Moodle material clickstreams and quiz scores and how to classify learning patterns that could potentially lead to obstacles to learning. In the proposed method, the deviation between the clickstream and the quiz score is calculated; this is then classified into four learning patterns based on the characteristics of those values and plus or minus signs. The four patterns classified here can be easily visualized through the creation of a scatter-plot graph. In addition, the scattered learning patterns increase the likelihood of identifying learners who should achieve excellent grades and those who are likely to fail the course. Here, the deviation is D_i, the observed value is x_i, and the average is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, calculated by the following equation: $D_i = x_i - \bar{x}$.

The first quiz was held in the third week of the fall semester of 2019. Table 1 (Week 3, Chapter 1, October 2, 2019) shows an example in which four learning patterns are classified from the deviation between the material clickstream and the quiz score at that time. In the work file in Table 1, the first column student shows the names of the anonymized learners, the second column Click shows the number of clickstreams of the materials in the chapter corresponding to the quiz range, the third column Quiz shows the quiz score, the fourth column Devi.click shows the deviation of the clickstream, the fifth column Devi.score shows the deviation of the quiz score, and the sixth column Learning Pattern shows the corresponding learning pattern. Patterns 1 to 4 correspond to the first to fourth quadrants on the scatter plot.

4.2 Learning Pattern by Deviation

As mentioned above, the learning pattern is considered to indicate behavioral engagement, so it is assumed that this affects the clickstream and the quiz score accumulated in the learning log of Moodle. Therefore, the number of outliers that appeared in the final quiz scores and the clickstream for the entire semester was examined. If a correlation exists between the number of occurrences of each learning pattern and the outliers, it is considered that the classified learning patterns can play a role similar to that of the outliers. This means that the visualization of learning patterns can help identify anomalous learners based on their final quiz scores and clickstreams, or learners who are struggling in class.

Table 1. *Example of the deviation calculation results and learning pattern classification for Week 3 (Chapter 1).*

	Click	Quiz	Devi.click	Devi.score	Learning Pattern
Student01	12	10	-17.09	2.72	P2
Student02	36	2	6.91	-5.28	P4
Student03	52	8	22.91	0.72	P1
Student04	22	10	-7.09	2.72	P2
Student05	24	8	-5.09	0.72	P2
Student06	11	6	-18.09	-1.28	P3
Student07	44	8	14.91	0.72	P1
Student08	21	4	-8.09	-3.28	P3
Student09	49	10	19.91	2.72	P1
Student10	15	8	-14.09	0.72	P2
Student11	32	4	2.91	-3.28	P4
Student12	27	8	-2.09	0.72	P2
Student13	24	8	-5.09	0.72	P2
Student14	27	8	-2.09	0.72	P2
Student15	22	4	-7.09	-3.28	P3
Student16	37	6	7.91	-1.28	P4
Student17	22	6	-7.09	-1.28	P3
Student18	15	2	-14.09	-5.28	P3
Student19	22	6	-7.09	-1.28	P3
Student20	20	8	-9.09	0.72	P2
Student21	24	6	-5.09	-1.28	P3
Student22	30	10	0.91	2.72	P1
Student23	21	8	-8.09	0.72	P2
Student24	29	6	-0.09	-1.28	P3
Student25	37	8	7.91	0.72	P1
Student26	35	8	5.91	0.72	P1
Student27	28	8	-1.09	0.72	P2
Student28	20	10	-9.09	2.72	P2
Student29	35	10	5.91	2.72	P1
Student30	31	6	1.91	-1.28	P4
Student31	27	8	-2.09	0.72	P2
Student32	29	10	-0.09	2.72	P2
Student33	31	8	1.91	0.72	P1
Student34	24	4	-5.09	-3.28	P3
Student35	47	10	17.91	2.72	P1
Student36	22	6	-7.09	-1.28	P3
Student37	32	8	2.91	0.72	P1
Student38	40	6	10.91	-1.28	P4
Student39	25	8	-4.09	0.72	P2
Student40	27	6	-2.09	-1.28	P3
Student41	40	10	10.91	2.72	P1
Student42	19	8	-10.09	0.72	P2
Student43	39	10	9.91	2.72	P1
Student44	34	4	4.91	-3.28	P4
Student45	24	8	-5.09	0.72	P2
Student46	49	4	19.91	-3.28	P4
Student47	34	10	4.91	2.72	P1
AVERAGE	29.085	7.277	0.000	0.000	
STDEV.S	9.717	2.223	9.717	2.223	
MAX	52	10	22.915	2.723	
MIN	11	2	-18.085	-5.277	
Data(N)	47	47	47	47	

Table 2. *Accumulation of learning patterns for weeks 1 to 15. "abs" = absent. Colored cells for click and quiz indicate outliers.*

	Clickstream(Week1-15)			Final Quiz	Learning Patterns					Hotelling's T2 theory			
	In class	Out of class	Total	Week 15	P1	P2	P3	P4	Abs	Outliers		P-value	
										Click	Quiz	Click	Quiz
Student01	316	2	318	22	0	8	5	0	0	2.972	0.001	0.005	1.000
Student02	412	128	540	21	4	5	4	0	0	0.029	0.039	0.977	0.969
Student03	268	26	294	15	0	1	9	0	3	3.580	2.324	0.001	0.024
Student04	442	133	575	27	2	2	4	2	3	0.006	1.275	0.995	0.208
Student05	477	63	540	24	1	5	1	3	3	0.029	0.217	0.977	0.829
Student06	478	49	527	20	4	6	1	1	1	0.068	0.175	0.946	0.862
Student07	354	194	548	24	5	3	4	0	1	0.013	0.217	0.990	0.829
Student08	521	205	726	27	7	2	1	3	0	1.285	1.275	0.205	0.208
Student09	330	51	381	21	0	2	9	0	2	1.645	0.039	0.107	0.969
Student10	521	148	669	22	6	2	0	4	1	0.539	0.001	0.592	1.000
Student11	254	33	287	21	0	4	4	0	5	3.767	0.039	0.000	0.969
Student12	568	189	757	22	7	2	0	3	1	1.824	0.001	0.074	1.000
Student13	447	111	558	26	4	3	4	2	0	0.002	0.825	0.999	0.414
Student14	419	253	672	27	8	2	2	1	0	0.571	1.275	0.571	0.208
Student15	360	116	476	24	1	5	4	1	2	0.381	0.217	0.705	0.829
Student16	389	4	393	14	0	3	8	2	0	1.437	3.048	0.157	0.004
Student17	550	45	595	24	5	4	3	1	0	0.047	0.217	0.963	0.829
Student18	302	41	343	17	0	5	7	0	1	2.399	1.171	0.020	0.247
Student19	339	111	450	22	0	3	7	1	2	0.639	0.001	0.526	1.000
Student20	490	218	709	21	2	3	2	5	1	1.029	0.039	0.309	0.969
Student21	412	22	434	10	0	0	11	2	0	0.831	6.919	0.410	0.000
Student22	469	159	628	22	8	2	2	1	0	0.200	0.001	0.842	1.000
Student23	540	31	571	17	1	3	5	4	0	0.002	1.171	0.998	0.247
Student24	407	69	476	24	2	4	4	1	2	0.381	0.217	0.705	0.829
Student25	466	250	716	19	9	0	0	4	0	1.131	0.410	0.264	0.684
Student26	404	16	420	23	3	7	3	0	0	1.019	0.060	0.313	0.953
Student27	453	49	502	24	2	8	3	0	0	0.189	0.217	0.851	0.829
Student28	393	147	540	22	0	7	4	2	0	0.029	0.001	0.977	1.000
Student29	432	181	613	22	7	3	1	2	0	0.117	0.001	0.907	1.000
Student30	497	111	608	28	4	3	4	2	0	0.094	1.824	0.925	0.075
Student31	408	37	445	9	0	2	11	0	0	0.696	8.131	0.490	0.000
Student32	424	391	815	20	4	2	1	3	3	3.086	0.175	0.003	0.862
Student33	572	187	759	27	10	1	0	2	0	1.862	1.275	0.069	0.208
Student34	432	139	571	16	0	1	11	1	0	0.002	1.699	0.998	0.096
Student35	662	211	873	24	12	0	1	0	0	4.679	0.217	0.000	0.829
Student36	460	76	536	18	2	3	6	2	0	0.039	0.742	0.969	0.462
Student37	494	59	553	25	3	4	3	0	3	0.006	0.472	0.995	0.639
Student38	562	154	716	28	8	2	0	3	0	1.131	1.824	0.264	0.075
Student39	399	60	459	25	1	7	5	0	0	0.542	0.472	0.590	0.639
Student40	361	170	530	21	1	5	6	1	0	0.057	0.039	0.955	0.969
Student41	480	199	679	27	11	2	0	0	0	0.647	1.275	0.521	0.208
Student42	344	30	374	25	0	8	3	0	2	1.773	0.472	0.083	0.639
Student43	438	231	669	28	11	1	0	1	0	0.539	1.824	0.592	0.075
Student44	474	55	529	16	4	2	4	3	0	0.061	1.699	0.952	0.096
Student45	558	61	619	17	2	6	2	3	0	0.148	1.171	0.883	0.247
Student46	574	270	844	26	11	0	0	1	1	3.841	0.825	0.000	0.414
Student47	490	187	678	25	7	2	2	2	0	0.636	0.472	0.528	0.639
AVERAGE	443.4	120.7	564.1	21.9	3.8	3.3	3.6	1.5	0.8	0.979	0.979	0.558	0.576
STDEV.S	86.7	86.1	142.8	4.5	3.7	2.2	3.1	1.4	1.2	1.214	1.581	0.382	0.373
MAX	662	391	873	28	12	8	11	5	5	4.679	8.131	0.999	1.000
MIN	254	2	287	9	0	0	0	0	0	0.002	0.001	0.000	0.000
0 (zero)					12	4	8	15	29				
Exclude 0					35	43	39	32	18				
Data(N)	47	47	47	47	47	47	47	47	47	47	47	47	47

Therefore, the number of occurrences of each learner from patterns 1 to 4 was aggregated from weeks 3 to 15, and correlations with the final quiz score and clickstream outliers were verified (Table 2, columns the eleventh to the fourteenth, Hotelling's T^2 theory). Table 2, columns the sixth to the tenth Learning Patterns shows the corresponding patterns 1 to 4 and the absenteeism data for each learner, as well as the calculated results for detecting outliers in clickstreams and quizzes. Specifically, the Pearson correlation coefficient was calculated using the data from columns the fourth to the ninth in Table 2. The detection of an outlier is based on Hotelling's T^2 theory, and the following equation is used, where the outlier is a, the observed value is $x_1, x_2, \dots, x_n$, the average is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and the standard deviation is $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$, calculated by the following equation: $\mathrm{a(x_i)} = (\mathrm{x_i}\text{-}\bar{\mathrm{x}})^2/\mathrm{s}^2$

4.3 Pearson's Correlation Coefficient between the Learning Log and the Four Patterns

In each learning pattern shown in Table 2, the data in which the number of occurrences is zero are included. Therefore, excluding those with zero occurrences, the Pearson's correlation coefficient was obtained for each learning pattern and final quiz score, as well as between each learning pattern and clickstream; the level of correlation was investigated. The following relationships for each pattern were clarified. Regarding absenteeism, the correlation coefficient was omitted because the number of data points was small.

Pattern 1 tends to exhibit both material clickstreams and quiz scores that are higher than average because of the opening of materials and reading them, and it is expected that the opening of materials will affect quiz scores. The Pearson's correlation coefficient between the final quiz scores of the learners who belonged to pattern 1 was $r = 0.396$ ($p = 0.009$, $p < 0.05$), indicating a weak positive correlation. The Pearson's correlation coefficient between the material clickstreams of the learners who belonged to pattern 1 was $r = 0.747$ ($p = 0.000$, $p < 0.05$), indicating a strong positive correlation. Learners who maintain this pattern at all times are expected to have good results and tackle their classes well. Another feature of this learning pattern is that it rarely exhibits pattern 3 (Table 2, students 33, 38, 41, 43, and 46).

Pattern 2 has a lower number of material clickstreams, but the quiz scores are higher than the average value. The Pearson's correlation coefficient between pattern 2 learners' final quiz scores and the pattern's number of occurrences was $r = 0.055$ ($p = 0.362$, $p > 0.05$), indicating no correlation. The Pearson's correlation coefficient between the clickstreams of material and the number of occurrences corresponding to the pattern was $r = -0.469$ ($p = 0.001$, $p < 0.05$), indicating a negative correlation. This means that learners who show pattern 2 tend to reduce the number of times they fall into this pattern if they open the material and read it carefully. Learners belonging to this pattern tend to have higher quiz scores without opening materials. Therefore, it is assumed that they have prior experience of learning content similar to that of the material; it is also assumed that some learners simply do not read the material. Furthermore, by listening carefully to the commentary while watching the teacher's monitor during class, some learners highly likely decided that they did not have to read the materials (Table 2, students 01, 09, 11, 19, 26, and 42).

Learners who exhibit pattern 3 have a relatively lower than average material clickstream and a lower than average quiz score. The Pearson's correlation coefficient between pattern 3 final quiz scores and the number of occurrences was $r = -0.689$ ($p = 0.000$, $p < 0.05$), indicating a negative correlation. The Pearson's correlation coefficient between the clickstream of the materials and the number of occurrences corresponding to the pattern was $r = -0.579$ ($p = 0.000$, $p < 0.05$), indicating a negative correlation. Learners who exhibit pattern 3 are assumed to have the tendency to be uninterested in the lesson content. It is expected that some of these learners will fall into the category of unsuccessful learners (Table 2, students 03, 16, 18, 21, and 31).

Learners who exhibit pattern 4 have more material clickstreams than average, but their quiz scores are lower than average. The Pearson's correlation coefficient between pattern 4 learners' final quiz scores and the group's number of occurrences was $r = -0.327$ ($p = 0.036$, $p < 0.05$), indicating a weak negative correlation. The Pearson's correlation coefficient between the clickstream of the materials and the number of occurrences corresponding to the pattern is $r = 0.328$ ($p = 0.033$, $p < 0.05$), indicating a weak positive correlation. This pattern includes learners who have not read the materials even though they have opened them and learners who cannot understand the materials even if they have read them. It is expected that those who are not interested in the class, as well as pattern 3 learners, fall under pattern 4. Giving priority to the learners who belong to this group and encouraging them to carefully read the materials and to concentrate and participate in the lessons are necessary (Table 2, students 20, 25, and 32).

5. Discussion

It is assumed that a correlation is established between learners' clickstream of the materials and their proper learning of the materials, as reflected by their weekly quiz and final quiz scores. When a positive correlation exists, it is assumed that the learner opened the materials and properly understood the lesson content. Conversely, there could be several causes for a negative correlation, such as inappropriate reading of the materials by the learners, insufficient understanding of the lesson content, use of inappropriate teaching methods, and inappropriate content of the materials.

However, during each lesson in this study, the course materials were displayed on the teacher's monitor screen, the content was explained to the learners as they read it, and the related Excel operations were demonstrated. Therefore, it is presumed that the learners—without having to open the materials on their own computer, often understood the lesson content from watching their teacher's explanations and demonstrations. This seemed to be true for learners who exhibited pattern 2. Therefore, if there are many learners who exhibit pattern 2, the relationship between the clickstream and the quiz score tends to be weakly correlated or uncorrelated.

6. Conclusion

A method was proposed to classify learning patterns using the deviation between the teaching material clickstream and the quiz score. The proposed method not only identified those learners who were struggling in the lesson but also showed the level of the learners' engagement in learning and their reaction to their teacher's teaching. It was found that few learners have a constant learning pattern, and many tend to change their learning patterns every week. However, learning patterns categorized from the Moodle course log can be used by teachers to identify and improve their approaches to better teaching and learning.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 18K11588 and 21K12183.

References

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 267-270).

Estacio, R. R., & Raga Jr, R. C. (2017). Analyzing students online learning behavior in blended courses using Moodle. Asian Association of Open Universities Journal, 12(1), 52-68

Gitinabard, N., Khoshnevisan, F., Lynch, C. F., & Wang, E. Y. (2018). Your actions or your associates? Predicting certification and dropout in MOOCs with behavioral and social features. *arXiv preprint arXiv:1809.00052*.

Hsiao, C. C., Huang, J. C., Huang, A. Y., Lu, O. H., Yin, C. J., & Yang, S. J. (2019). Exploring the effects of online learning behaviors on short-term and long-term learning outcomes in flipped classrooms. *Interactive Learning Environments*, *27*(8), 1160-1177.

McKay, T., Miller, K., & Tritz, J. (2012). What to do with actionable intelligence: E2Coach as an intervention engine. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 88-91).

Mouri, K., Ren, Z., Uosaki, N., & Yin, C. (2019). Analyzing learning patterns based on log data from digital textbooks. *International Journal of Distance Education Technologies (IJDET)*, *17*(1), 1-14.

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for educational data mining: A review. Journal of Educational and Behavioral Statistics, 42(1), 85-106.

Steinmayr, R., Weidinger, A. F., Schwinger, M., & Spinath, B. (2019). The importance of students' motivation for their academic achievement–replicating and extending previous findings. *Frontiers in psychology, 10*, 1730.

Skinner, E. A., Kindermann, T. A., & Furrer, C. J. (2009). A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and psychological measurement*, *69*(3), 493-525.

Vermunt, J. D., & Donche, V. (2017). A learning patterns perspective on student learning in higher education: state of the art and moving forward. *Educational psychology review*, *29*(2), 269-299.

Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Identifying Learner Problems Framed within MOOC Learning Designs

Paraskevi TOPALI*, Alejandro ORTEGA-ARRANZ*, Alejandra MARTÍNEZ-MONÉS, Sara VILLAGRA-SOBRINO, Juan I. ASENSIO-PÉREZ & Yannis DIMITRIADIS
GSIC-EMIC Research Group, Universidad de Valladolid, Spain
*{evi.topali | alex}@gsic.uva.es

Abstract: Detecting learners who face problems in MOOCs usually poses difficulties due to the high instructor-learners ratio, the diversity of the population, and the asynchronous participation mode. Existing solutions mainly draw on self-reported problems in discussion forums and on dashboards displaying learners' activity traces. However, these approaches cannot scale up easily or do not consider the course learning design. This paper presents a conceptual framework aimed at guiding MOOC instructors in the identification of potential learners' problems and indicators of such problems, considering the learning design of the course (e.g., types of activities, difficulty, etc.). An instrumental qualitative case study served for the evaluation and refinement of the framework. The results showed that the framework positively helped instructors to reflect on potential learners' problems they had not considered beforehand, and to associate such problems with a set of indicators related to their learning designs.

Keywords: MOOCs, Instructors, Design Framework, Learner Identification, Co-design

1. Introduction

MOOCs have been prominent in the educational domain and, during the last year, with the coronavirus outbreak, they have experienced a drastic growth in the number of new users and courses provided with respect to the last previous years (Shah, 2020). However, despite this high number of users, support to learners during MOOC enactment has been rather overlooked (Gregori, Zhang, Galván-Fernández, & Fernández-Navarro, 2018), still presenting research challenges such as the provision of timely and useful feedback for those learners experiencing course-related problems (Aldowah, Al-Samarraie, Alzahrani, & Alalwan, 2020). Current practices to identify learners who face problems in MOOCs, as well as to assist them, regularly require that instructors look at posts explicitly reporting these problems in discussion forums (Onah, Sinclair, & Boyatt, 2014). Yet, the high number of posts (Shatnawi, Gaber, & Cocea, 2014), the diversity regarding learners' knowledge (Deboer, Seaton, & Breslow, 2013) and the asynchronous learners-instructor communication, pose doubts about its scalability for instructors.

To that end, the use of Learning Analytics (LA) is explored to automate the identification of learners' problems and the provision of personalized feedback (Gašević, Dawson, Rogers, & Gasevic, 2016). Mainstream MOOC platforms (e.g., Coursera, Open edX, Canvas Network) usually store data regarding participants' behavioural footprints generated at course runtime (Jansen, van Leeuwen, Janssen, & Kester, 2020), such as the interactions with other learners (e.g., number of posts), the interactions with course resources (e.g., PDF downloads, video views), or the student learning outcomes (e.g., quiz scores, attempts). This information can be displayed through dashboards to keep learners aware of their performance, and to assist instructors in the detection of critical learner behaviours (Urrutia, Cobos, Dickens, White, & Davis, 2016). However, current LA strategies used for identifying learners who may need further assistance have not been founded on pedagogical strategies for instruction (Gašević, Dawson, & Siemens, 2015). Concretely, the information displayed in the dashboards does not usually consider the course characteristics or the pedagogical intentions of the instructors. For instance, the work by Dabbebi et al., (2017) revealed that, in the case of a dashboard, not all collected student data were equally meaningful for MOOC instructors, since the learning context determines whether some data is more or less informative than other.

Gašević et al., (2016) argued that LA should be informed by the course context and learning design (LD) to result in useful conclusions and appropriate intervention. One approach to get this information is to explicitly obtain it from MOOC instructors, by including them in the loop and making them actors of the decision-making process (Rodríguez-Triana, Prieto, Martínez-Monés, Asensio-Pérez, & Dimitriadis, 2018). This same approach can be applied to the design of detection strategies for learners facing problems. This way, instructors would be involved in how to identify learners with problems and how to assist them, based on their expertise. For instance, instructors are the ones aware of the difficulty of the activities, the pedagogical connections between the different course activities, or the relevant indicators that might point out problematic situations.

Given this context, the identification of potential problems that learners might face during the MOOC, as well as the identification of indicators that might help in the detection of learners facing them, are two crucial aspects which should be considered to shape useful feedback interventions. To the best of our knowledge, previous solutions did not consider the LD information to detect learners facing problems and do not guide instructors in such a process. In this paper, we present a study that aims to answer the following research question (RQ): ***How can instructors be supported in the identification of potential learners' problems considering LD parameters at design time?*** To answer the RQ, we proposed a conceptual framework, named FeeD4Mi, to help MOOC instructors in the reflection and identification of problems and indicators pointing to potential learners' difficulties within their LDs. Consequently, two sub-research questions associated to the previous RQ were defined:

a. *RQ-1: To what extent did FeeD4Mi cover the problems and indicators potentially describing learners' difficulties within MOOC LDs?*
b. *RQ-2: To what extent did FeeD4Mi facilitate instructors to reflect on additional problems and indicators for MOOC LDs?*

The structure of the paper is as follows. Section 2 introduces the proposed framework. Next, Section 3 describes the design of the study including the context, the participants, and the data sources. Finally, the results are presented (Section 4) and discussed together with ideas for future work (Section 5).

2. Framework Presentation

FeeD4Mi is a five-dimension framework foreseen to be employed during the design phase of the course and expected to facilitate MOOC instructors to: (a) recognize potential problems in MOOCs; (b) define potential behaviours of learners having an expected problem; and (c) choose the adequate support for the identified problems. We defined FeeD4Mi through a research process, based on a literature review (e.g., Aldowah et al., 2020; Botelho et al., 2019; Henderikx, Kreijns, & Kalz, 2018) and on experiences with MOOC instructors (Topali et al., 2019; Topali, Ortega-Arranz, Martínez-Monés, & Villagrá-Sobrino, 2020), regarding the detection of learners facing problems, from which we abstracted the important parameters identified in the provision of feedback practices. The final proposal encompasses five dimensions (see Figure 1) and a sequence of steps, related to each dimension, as described below:

- **Learning Design Analysis.** Learners' problems and feedback practices are context dependent. LDs contain information useful for the design of feedback regarding activities' objectives, tools, resources and expected outcomes (Gašević et al., 2016). We can derive such information from the instructors, who are the ones responsible of the pedagogical strategies applied. That is, the consideration of the instructors' course decisions (e.g., difficulty of the activities), the course components, and the connection among such components (e.g., the answers to Quiz A are given in Video I and Video II).
- **Reflection on Learner Problems.** This step encompasses a reflection on the learners' problems that can occur in a MOOC during the learning process. FeeD4Mi includes a catalogue of problems regarding content-related, peer collaboration, technical, learners' assistance, community building and self-regulation issues.
- **Selection of Problem Indicators.** This step deals with the detection of indicators that point to behaviours associated with the identified problems. These indicators can be classified into self-reported (e.g., private messages referring to the problem) and log data (e.g. number of attempts in assignments, time spent in the course, etc.) indicators.
- **Reflection on Feedback Conditions and on Feedback Aspects.** These two steps involve the creation of rule-based decisions and the design of feedback interventions based on the last two

dimensions of FeeD4Mi (see Figure 1). These dimensions are guided by the factors of feedback quality described by Molloy & Boud (2014). According to such factors, when designing feedback, educators need to define: a) the feedback provider (e.g., instructor, peers, context), b) the time (e.g., on time feedback or delayed on purpose), and c) the feedback type (e.g., hints or direct feedback). FeeD4Mi includes a catalogue of scalable feedback practices.

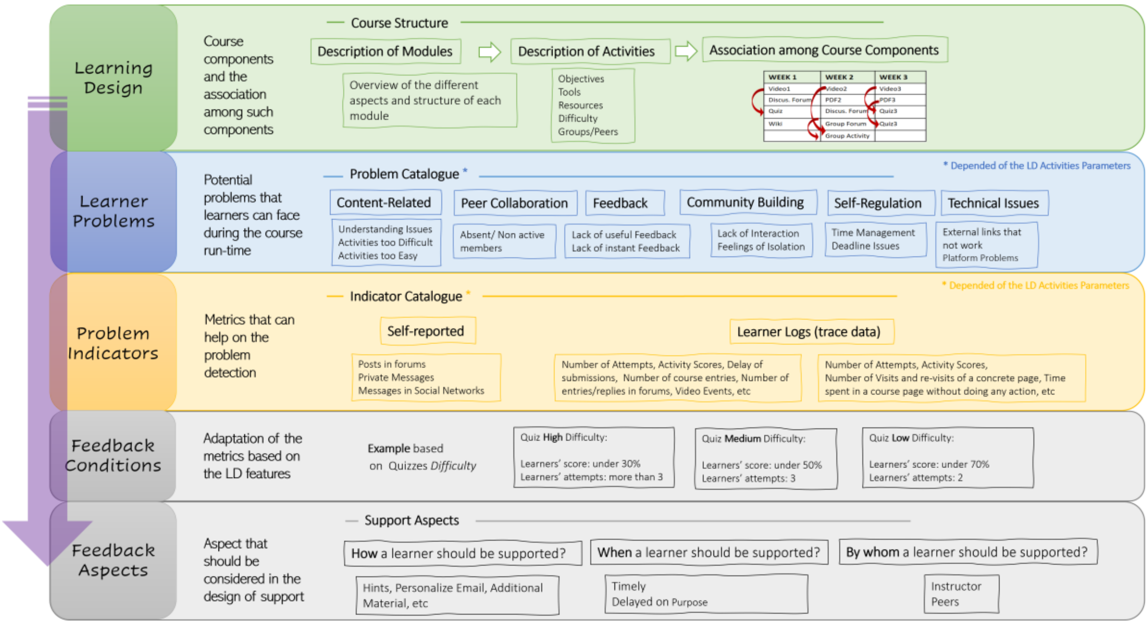

Learning Design
Course components and the association among such components
Course Structure
Description of Modules
Description of Activities
Association among Course Components
Overview of the different aspects and structure of each module
Objectives
Tools
Resources
Difficulty
Groups/Peers
WEEK 1
WEEK 2
WEEK 3
Video1
Discus. Forum
Quiz
Wiki
Video2
PDF2
Discus. Forum
Group Forum
Group Activity
Video3
PDF3
Quiz3
Quiz3
Learner Problems
Potential problems that learners can face during the course run-time
Problem Catalogue *
* Depended of the LD Activities Parameters
Content-Related
Understanding Issues
Activities too Difficult
Activities too Easy
Peer Collaboration
Absent/ Non active members
Feedback
Lack of useful Feedback
Lack of instant Feedback
Community Building
Lack of Interaction
Feelings of Isolation
Self-Regulation
Time Management
Deadline Issues
Technical Issues
External links that not work
Platform Problems
Problem Indicators
Metrics that can help on the problem detection
Indicator Catalogue *
* Depended of the LD Activities Parameters
Self-reported
Posts in forums
Private Messages
Messages in Social Networks
Learner Logs (trace data)
Number of Attempts, Activity Scores, Delay of submissions, Number of course entries, Number of entries/replies in forums, Video Events, etc
Number of Attempts, Activity Scores, Number of Visits and re-visits of a concrete page, Time spent in a course page without doing any action, etc
Feedback Conditions
Adaptation of the metrics based on the LD features
Example based on Quizzes Difficulty
Quiz High Difficulty: Learners' score: under 30% Learners' attempts: more than 3
Quiz Medium Difficulty: Learners' score: under 50% Learners' attempts: 3
Quiz Low Difficulty: Learners' score: under 70% Learners' attempts: 2
Feedback Aspects
Aspect that should be considered in the design of support
Support Aspects
How a learner should be supported?
Hints, Personalize Email, Additional Material, etc
When a learner should be supported?
Timely
Delayed on Purpose
By whom a learner should be supported?
Instructor
Peers

Figure 1. FeeD4Mi Overview Presenting the Five Dimensions and Their Content.

3. Methodology

In this section, we report an evaluation of FeeD4Mi regarding the posed RQs (see Section 1). More concretely, the evaluation was designed as an instrumental qualitative case study (Creswell, 2006), consisting of two co-design sessions (Case#A and Case#B) with three MOOC instructors (two were instructors of the same course) who were preparing their upcoming MOOCs. Case#A was about a MOOC on the subject of EU-Russia Relations and was offered in Estonia. Case#B involved a MOOC about Programming and was offered in Greece. Both included different types of activities. The participants were selected following a purposive sample approach (Fraenkel, Wallen, & Hyun, 2012). That is, participants were selected due to their previous experience as MOOC instructors (on average, 3 MOOCs), and due to their intention to provide a MOOC in the upcoming weeks. During the co-design sessions, we requested participants to perform tasks associated to the first three FeeD4Mi dimensions:

1. **Summarize the course LD**: We asked participants to outline their course, describing the modules, the activities (e.g., quizzes, documents), and their relationships and features (e.g., difficulty, group activities), according to the first dimension of the framework.
2. **Reflect and specify learners' problems**: We asked participants to specify potential problems that learners could face in their courses based on the previous outline. Initially, participants brainstormed about problems without the support of FeeD4Mi. The problems mentioned allowed us to evaluate the *"completeness"* of the catalogue of problems included in FeeD4Mi (*RQ-1*). In a second step, we introduced participants to the problems enlisted in such catalogue according to the course LD. This allowed us to test the *"discoverability"* of the problems suggested by FeeD4Mi (*RQ-2*).
3. **Reflect and select problem indicators**: We asked participants to connect the mentioned problems with indicators that could help identify such problems (e.g., video metrics). Participants brainstormed on the indicators without the support of FeeD4Mi and we evaluated the *"completeness"* of the indicators' catalogue (*RQ-1*). Later, we presented the FeeD4Mi catalogue to help participants to reflect on additional indicators that might result informative, thus testing *"discoverability"* (*RQ-2*).

The data sources used in this evaluation were: the participants' artefacts (i.e., participants products created during the co-design experiences) *[Art_CaseX]*, which were analysed considering the

FeeD4Mi catalogues; the recordings of the sessions *[Rec_CaseX]*, from which the time employed was also measured; and the observations made by the leading researcher during such sessions *[Obs_CaseX]*.

4. Results

The analysis of the artefacts created by instructors revealed a total number of 9 potential problems that were identified without using FeeD4Mi (see white colour in Table 2). While instructors from Case#A focused more on content-related problems (e.g., difficulty of quizzes, academical writing in assignments), instructor of Case#B focused more on peer interaction problems such as communications in discussion forums and peer assessments. FeeD4Mi already included 5 of those problems (55.56%). Problems related to learners' familiarity with the course platform, learners' different backgrounds, and lack of proper interaction among peers were mentioned by participants and they were not included in FeeD4Mi (see '*' in Table 2). After being exposed to the FeeD4Mi catalogue of problems, participants considered additional issues that might be potentially relevant to their courses. Concretely, 2 potential problems in Case#A (25% additional) and 4 potential problems in Case#B (57.14% additional). For 2 of such problems, participants expressed their concern regarding the difficulty and unawareness of how to deal with such challenges (before being introduced to the FeeD4Mi catalogue of indicators).

In the task *Reflect and select problem indicators*, participants identified 19 different indicators that may provide alerts on the previous problems (see Table 2). The FeeD4Mi catalogue of indicators already considered 14 of them (73.68%). It seems interesting to highlight that all the non-included indicators require content analysis for their interpretation, such as the analysis of the learners' submitted work or the content of forum posts. In the second step of this task using FeeD4Mi, a total number of 6 indicators were pointed out as useful (20.69%). Additionally, as expressed by Case#A instructors, while some indicators may not be meaningful enough alone, their combination with other indicators could reveal potential problems (e.g., the time spent in a quiz together with the number of video watches).

Furthermore, we evaluated also the effort associated to the sequential process related with the use of FeeD4Mi (which was performed during the co-design sessions). To that end, we analysed the data sources to understand the suitability and the difficulties of such process within instructors' regular MOOC practice (see Table 3). The excerpts suggest that the whole process helped participants to further reflect about potential learners' problems and to specify their own LD (see Table 3, *Positive*). Nevertheless, it is worth mentioning that participants also reported emerging negative impressions from such a process. Specifically, the reflection on problem indicators seemed a complex task for them that required extra effort (see Table 3, *Negative*). Additionally, the long duration of the process, on average 1.5 hours, was considered tiring for the instructors, who at the end, wanted to quickly finish the session.

5. Discussion & Conclusions

This study focused on supporting MOOC instructors in the identification of potential learners' problems and indicators, that may provide alerts on such problems, considering the course LD. To this end, we propose a conceptual framework, FeeD4Mi. *RQ-1* aimed at understanding the extent to which FeeD4Mi supports the problems and indicators associated with the LD as described by instructors. Results from the co-design sessions revealed that FeeD4Mi directly supports 55.56% and 73.68% of the problems and indicators, respectively. All non-supported indicators require content analysis of learners' posts and artefacts. This evaluation provided useful insights to complement the current catalogue of problems, although further work is needed to investigate useful indicators for such new potential problems.

RQ-2 deals with the discoverability of FeeD4Mi to help instructors reflect and identify problems and indicators not considered before. Results showed that instructors identified 6 additional problems thanks to the reflection triggered by FeeD4Mi. Specially, the LD and the association of problems with the different components helped to detect tricky course parts that can be challenging for learners. Also, it contributed in improving specific course aspects, such as the type and nature of activities. Moreover, instructors identified 6 additional indicators from the FeeD4Mi catalogue.

The co-design sessions also revealed that using FeeD4Mi was not a trivial task for MOOC instructors. In practice, we observed that the process associated to FeeD4Mi seemed time consuming and complex, especially towards the reflection on indicators. Likely, such complexity was influenced by the fact that the process lasted 1.5 hours. Consequently, it seems interesting to explore whether the time needed, and the complexity of the process can be reduced, and if these results are also transferable to novice MOOC instructors. The evaluation allowed us to collect initial evidence of the FeeD4Mi benefits and insights to refine the process and catalogues. This study presents some limitations as it is based on two co-design sessions involving only three MOOC instructors. As a future work, we plan to perform an evaluation with more instructors to understand the extent to which the results obtained in this study can be generalized for multiple instructors and course topics. This evaluation could also incorporate the remaining dimensions of the framework, aiming at a comprehensive overview of the framework benefits for the creation of instructor-designed feedback strategies in MOOCs.

Table 2. *Identified learners' problems and indicators. Grey: Additional aspects emerged from the reflection with FeeD4Mi. *Aspects reported by the participants which were not included in FeeD4Mi*

	Problem	Association with the LD	Problems' Indicators
Case#A: EU-Russia Relations	Misunderstanding of the given task	Discussion forums and quizzes of modules 1-4	Posts in discussion forums Email from the learners A lot of time spent in a page
	Issues of academic writing and referencing (various levels of knowledge)	Assignments of modules 4-5	*Analyzing submitted work Email from the learners More attempts in a quiz
	*Communication skills	Discussion Forums	Post Interaction: entries and replies
	*Students are not familiar with LMS functions	Module 0	Email from the learners *Posts in wrong spaces *Check post-course survey
	Issues of connectivity and accessibility of various interactive materials	Content page, content videos and videos recap.	Logs of course access Check technical questions Email from the learners
	*Language issues	Whole Course	Video features (pause, forward)
	Absent/ non-active members	Whole Course	Check post-course survey
	Deadline / Time issues	Whole Course	Posts in discussion forums Email from the learners Delays of activity submissions
Case#B: Programming	Peer assessment	Projects of modules 1-5	Scores in peer feedback
	*Different backgrounds	Discussion Forums	*Naïve/advanced questions
	Low participation forums	Discussion Forums	A lot of visits in the forums
	Lack of instant feedback	Discussion Forums and Emails	Posts in discussion forums Non replies in posts / emails
	Understanding / Content issues	Whole Course	Posts in discussion forums Scores in quizzes
	Activities too difficult	Whole Course	Posts in discussion forums Scores in quizzes under thresholds
	Deadline / Time issues	Projects of modules 1-5	Posts in discussion forums Many posts of the same problem

Table 3. *Excerpts Related with the Co-Design Process.*

Categories	Labels	Excerpts of Evidence
Positive	*[Rec#CaseA]*	*"I think it was useful to reflect on the things that we should maybe pay attention to. [..]I think that for future planning, it's also relevant".*
	[Rec#CaseB]	*"I haven't noted the course design and what we created is really useful".*
Negative	*[Obs#CaseB]*	*The identification of indicators is more challenging to proceed than the identification of the problems who run more smoothly.*
	[Rec#CaseA]	*"I feel we are not very creative with our indicators".*

Acknowledgements

The authors would like to thank Prof. Nikolaos Avouris, Prof. Anna Beitane, and Ms. Mariam Tlashadze for participating in this study, which has been funded by the Spanish State Research Agency and the European Regional Development Fund (TIN2017-85179-C3-2-R, PID2020-112584RB-C32); by the European Regional Development Fund and the Regional Government of Castilla y León (VA257P18) and the European Social Fund and Regional Council of Education of Castilla y León.

References

Aldowah, H., Al-Samarraie, H., Alzahrani, A. I., & Alalwan, N. (2020). Factors affecting student dropout in MOOCs: A cause and effect decision- making model. *J. of Computing in Higher Education*, *32*(2), 429–454.

Botelho, A., Varatharaj, A., Patikorn, T., Doherty, D., Adjei, S., & Beck, J. (2019). Developing early detectors of student attrition and wheel spinning using deep learning. *IEEE Trans. Learning Techn.*, *12*(2), 158–170.

Creswell, J. (2006). Five qualitative approaches to inquiry. In *Qualitative inquiry and research design*, pp. 53–84. Sage Publications, Inc.

Dabbebi, I., Iksal, S., Gilliot, J.-M., May, M., & Garlatti, S. (2017). Towards adaptive dashboards for learning analytic: An approach for conceptual design and implementation. In *Proc. of the 9th Int. Conf. on Computer Supported Education*, (pp. 120–131).

Deboer, J., Seaton, D. T., & Breslow, L. (2013). Diversity in MOOC students backgrounds and behaviors in Relationship to performance in 6.002x. In *Proc. of 6th Learning Int. Networks Consortium Conf.*, (pp. 1–10).

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education*. McGraw-Hill Education.

Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *Internet and Higher Educ. 28,* 68–84.

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, *59*, 64–71.

Gregori, E. B., Zhang, J., Galván-Fernández, C., & Fernández-Navarro, F. (2018). Learner support in MOOCs: Identifying variables linked to completion. *Computers and Education*, *122*, 153–168.

Henderikx, M., Kreijns, K., & Kalz, M. (2018). A classification of barriers that influence intention achievement in MOOCs. In *Proc. of the 13th European Conference on Technology Enhanced Learning*, (pp. 3-15).

Jansen, R. S., van Leeuwen, A., Janssen, J., & Kester, L. (2020). A mixed method approach to studying self-regulated learning in MOOCs: Combining trace data with interviews. *Frontline Learning Research*, *8*(2), 35–64.

Molloy, E. K., & Boud, D. (2014). Feedback models for learning, teaching and performance. In *Handbook of Research on Educational Communications and Technology: Fourth Edition* (pp. 413–424).

Onah, D., Sinclair, J., & Boyatt, R. (2014). Exploring the use of MOOC discussion forums. In *Proc. of London Int. Conf. on Education*, (pp. 1–4).

Rodríguez-Triana, M. J., Prieto, L. P., Martínez-Monés, A., Asensio-Pérez, J. I., & Dimitriadis, Y. (2018). The teacher in the loop: Customizing multimodal learning analytics for blended learning. In *Proc. of the 8th Int. Conf. on Learning Analytics and Knowledge*, (pp. 417–426).

Shah, D. (2020). The Second Year of The MOOC: A review of MOOC stats and trends in 2020. Retrieved from https://www.classcentral.com/report/the-second-year-of-the-mooc/, last access: July 2021.

Shatnawi, S., Gaber, M. M., & Cocea, M. (2014). Automatic content related feedback for MOOCs based on course domain ontology. In *Int. Conf. on Intelligent Data Eng. and Automated Learning*, (pp. 27–35).

Topali, P., Ortega-Arranz, A., Er, E., Martínez-Monés, A., Villagrá-Sobrino, S. L., & Dimitriadis, Y. (2019). Exploring the problems experienced by learners in a MOOC implementing active learning pedagogies. In *Proc. of the 6th European MOOCs Stakeholders Summit,* (pp. 81–90)

Topali, P., Ortega-Arranz, A., Martínez-Monés, A., & Villagrá-Sobrino, S. L. (2020). "Houston, we have a problem": Revealing MOOC practitioners' experiences regarding feedback provision to learners facing difficulties. *Computer Applications in Engineering Education*, *29*(4), 769-785.

Urrutia, M. L., Cobos, R., Dickens, K., White, S., & Davis, H. (2016). Visualising the MOOC experience: a dynamic MOOC dashboard built through institutional collaboration. In *Proc. of the 4th European MOOCs Stakeholders Summit*, (pp. 1–8).

Investigating the Tightness of Connection between Original Map and Additional Map in Extension Concept Mapping

Didik Dwi PRASETYA[a,b*], Aryo PINANDITO[a,c], Yusuke HAYASHI[a] & Tsukasa HIRASHIMA[a]
[a]*Department of Information Engineering, Hiroshima University, Japan*
[b]*Department of Electrical Engineering, Universitas Negeri Malang, Indonesia*
[c]*Information System Department, Universitas Brawijaya, Indonesia*
*didikdwi@um.ac.id

Abstract: Extension concept mapping extends an existing original map by linking it to a new additional map. This technique encourages learners to review and improve their knowledge structure. Previous studies have demonstrated the difference of knowledge structure achievements between the original and additional maps in the Extended Scratch-Build (ESB) and Extended Kit-Build (EKB) approaches. However, no information has been provided related to the extent of the tightness between the concept maps. The tight interconnectedness of knowledge structures represents expertise and depth of personal knowledge. This study investigated the effect of different concept mapping tools on the student's ability to connect concept maps. Fifty-five second-year university students participated and were divided into two groups: the control group utilized the ESB map, and the experimental group used the EKB map. Extension Relationships (ER) scores were used to confirm that learners could associate prior existing with new concept maps. ER is a particular link that tightly interconnected the previous original map with the additional map. ER scores evaluate both the quantity and quality of relations that the learners have made. The statistical analysis results emphasized that the experimental group outperformed the control group regarding the number and quality of ER scores.

Keywords: Concept map, extension concept mapping, tight connection, extension relationships (ER)

1. Introduction

A concept map is a visual representation of an individual's knowledge structure. It explicitly expresses the most relevant relationships between a set of ideas (Novak, 2010). Conceptual knowledge in a particular domain requires a highly integrated structure to be meaningful and valuable in its context (Ruiz et al., 1997). In practice, Vanides et al. (2005) said, "highly proficient learners tend to produce highly interconnected maps, whereas novices often create simple structures." Hence, concept interrelatedness is an essential property of personal knowledge. Knowledge structures that are tightly interconnected encourage learners to recall information and achieve meaningful learning (Turns et al., 2000).

The construction style of a concept map can be categorized into two kinds: open-ended and closed-ended maps (Taricani and Clariana, 2006; Ruiz-Primo et al., 2001; Hirashima, 2019). Open-ended fashion is a mapping activity that allows learners to add links, concepts, and form propositions freely that express their knowledge. It enables the teacher to reveal the difference between students' knowledge structures (Ruiz-Primo et al., 2001; Hirashima, 2019). However, it is hard to assess (Taricani and Clarina, 2006) and provides feedback to learners. On the contrary, a closed-ended fashion provides finite links and concepts. The existence of map components enables individuals to recall critical concepts they have learned and reach the maximum test scores (Vanides et al., 2005). However, it is not easy to reveal the differences among students in the closed-ended method.

The initial study proposed Extended Kit-Build (EKB) concept mapping approach to reflect students' knowledge structure in the closed-ended method (Prasetya et al., 2019). The EKB employs a recomposition Kit-Build (KB) framework in providing a solid knowledge structure. A recomposition map is an important learning activity that requests learners to understand other's understanding. During the reconstruction, KB explicitly directs learners to understand the teacher's understanding (Yamasaki et al., 2010; Hirashima et al., 2015). Further investigations were carried out by comparing the performance of EKB and Extended Scratch-Build (ESB), a similar extension concept mapping design (Prasetya et al., 2021). ESB uses the open-ended technique to build the original map and extend it to produce the additional map (Prasetya et al., 2020). The results reported that although the size of the ESB's original map was broader, the overall achievement of the EKB's map was superior. However, there is no information available regarding the extent of the tightness between the original map and the additional map that describes an individual's meaningful learning achievement. Vanides et al. (2005) emphasized that highly interconnected concept maps represent proficiency and depth of personal knowledge.

The present study compared ESB and EKB concept mapping to reveal which approach promotes highly interconnected knowledge structure the extent of the effect. In addition, students' performance in both groups was measured using Extension Relationships (ER) scores consisting of the number of ER and quality of ER. The following research questions guide this study:

1. What is the effect of ESB and EKB approaches on the ability of students to connect concept maps tightly?
2. To what extent is the quality of ER in the EKB compared to the ESB concept mapping?

2. Literature Review

2.1 Extension Concept Mapping

Extension concept mapping is an activity for extending an existing concept map by integrating new relevant information. It provides learners with the opportunity to review initial ideas and connections, elicit missing ideas and relationships, adding new concepts and links, and revising knowledge integration (Foley et al., 2018; Schwendimann & Linn, 2016). Extension concept mapping comprises two interrelated activities creating an *original map* and an *additional map* (Prasetya et al., 2021).

The current study investigated the EKB concept mapping that employs Kit-Build (KB) framework to facilitate students in expressing their understanding (Prasetya et al., 2019). First, EKB's students were requested to recompose the teacher's map related to the original material. Next, they were asked to extend their concept map by adding new concepts and links. The EKB was compared with the ESB concept mapping to confirm its learning effects. ESB's students were allowed to add any concepts and any links in their blank canvas. Furthermore, they were requested to extend the previous map with the same technique. The results confirmed that the EKB group outperformed the ESB group regarding comprehension scores and map size.

2.2 Extension Relationships

Extension concept mapping connects the original map and additional map to form a unified knowledge structure. The relation that links tightly between the original map and the additional map can be called an extension relationship (ER). ERs are an essential component of extension concept mapping activities. The tightness of connection depicts enhanced meaningful learning and depth of personal knowledge (Vanides et al., 2005). In principle, ER is a proposition as in standard concept maps. However, this proposition links directly between the original map and the additional map. An illustration of the ER on the extension concept map can be shown in Figure 1.

Since the ER is a proposition, it could be assessed using proposition-based measurement. ER could be considered to confirm improved meaningful learning in extension concept mapping situations. ER measurement is based on the meaningful learning theory in which learners engage to make sense of their experience, connect one idea to another, and deliver what was studied to answer new obstacles. It denotes that learners are not merely able to apply previously acquired knowledge to build concept maps but also improve meaningful learning to attain new knowledge and solve complex problems. ER represents the deep and broadness of personal knowledge structure and could be assessed quantitatively and qualitatively.

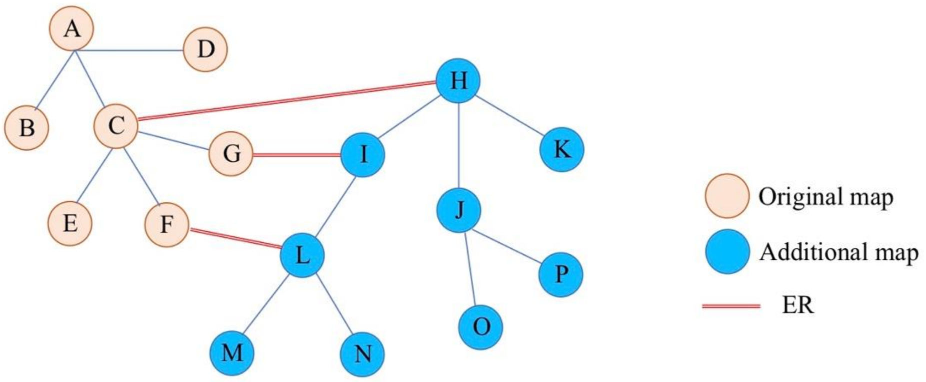


Figure 1. ER illustration in Extension Concept Mapping.

3. Methods

3.1 Participant and Context Material

The participants of this study were 55 second-year university students from two regular classes (A and B). A pre-test was conducted before determining the role of the group, and the results stated that both classes were homogeneous (p = .389 > .05). Therefore, class A was randomly assigned as the control group, and class B was the experimental group. The control group consisted of 27 participants, and the experimental group had 28 participants.

This study was conducted on the Database 1 course with Relational Database topic. The lecturer used a presentation, and she delivered paper-based handouts to students before the teaching activity. In the EKB approach, the lecturer first determined the teacher's map that will be decomposed and then given to students for reconstruction. Original material in the Relational Database topic consists of 283 words (10 slides), and the lecturer provided ten propositions. The content of the additional part consists of 237 words (8 slides), and there was no teacher's map provided for this phase. A senior lecturer who has 11 years of teaching experience taught in both groups.

3.2 Instruments

ER achievement was confirmed using two measurements: the number of ER and the quality of ER. The ER number was calculated through links directly related to the concept between the original and additional maps. The number of ERs represents the number of propositions that act as intermediaries on mapping activities. The amount of propositions symbolizes the broadness of student knowledge in a particular domain (Stoddart, 2006; Jaafarpour et al., 2016). ER demonstrates the extent of students' knowledge structure in extended concept mapping activities.

The ER quality was evaluated using the quality of propositions scores. The quality of propositions is one of the most critical and recommended judgments in the concept mapping assessment and states the quality of personal knowledge (Raud et al., 2016). The lecturer formulated the quality of the ER scoring method to examine the level of students' understanding. Four level scoring were defined: 0 = incorrect; 1 = partially incorrect; 2 = correct with thin scientific understanding; and 3 = scientifically correct. The quality measurements on both groups were judged manually by the same lecturer.

3.3 Procedures

Figure 2 depicts the experimental procedures of the control and experimental groups. Experimental activities begin by giving the original material to the learners for 25 minutes. The lecturer used the same approach in both groups, which consists of conveying the material and discussion as usual. After delivering the first material, next is the concept map construction of Phase 1. Students in the control group requested to create an original concept map used an open-ended technique, while those in the experimental group were utilized the KB approach. During the activity, students were allowed to read the handouts.

Furthermore, the teacher continued presenting additional material for both groups in Phase 2. The second teaching was carried out for 25 minutes with the same approach. Next, students in the two groups were requested to extend their previous map using the same approach within 15 minutes. Students in both groups could add any concepts, links, and form propositions related to the additional material and connect them to the original map. Students could review their previous maps in expanding maps, elicit missing elements, add new ideas, and improve knowledge structures.

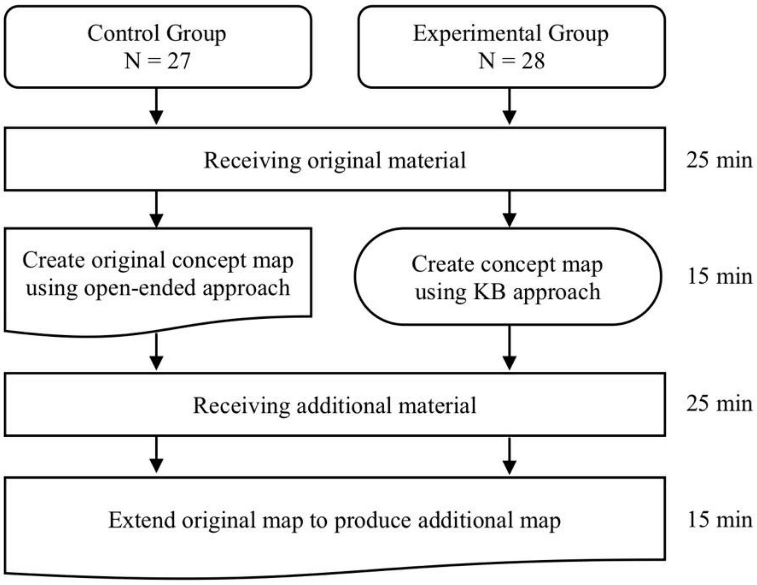

Control Group
N = 27
Experimental Group
N = 28
Receiving original material
25 min
Create original concept map using open-ended approach
Create concept map using KB approach
15 min
Receiving additional material
25 min
Extend original map to produce additional map
15 min

Figure 2. Experimental Procedure.

4. Results and Discussion

4.1 Analysis of the Number of ER

Detailed statistics on the achievement of ER numbers for both groups are shown in Table 1. The average achievement of students in the experimental group was higher than the control group, which was 5.32 compared to 1.63.

Table 1. *Descriptive Statistics of the Number of ER for Both Groups*

Group	N	Min	Max	Median	Mean	SD
Control group	27	1	2	2.00	1.63	0.49
Experimental group	28	2	11	4.00	5.32	3.10

Students' achievement on the ER number between the control and experimental groups was further analyzed using the Mann-Whitney U test. The results showed greatly significant difference for both groups (Z = -5.994; p = .000 < .05). The mean rank of the ER number for the control group was 1.63, while the experimental group was 5.32. The effect size, expressed through Pearson's r, could be calculated using Z values obtained from the Mann-Whitney U test. The results indicated a large effect size (Pearson's r = -.808).

Students' performance in the experimental group who used the EKB approach was superior to participants in the control group who used ESB. The EKB approach that employs KB recomposition on original map reconstruction impacts achieving the number of ERs. Recomposition teacher's map encourages learners to understand knowledge target well. It states an essential learning activity in KB that provides a solid knowledge structure to promote meaningful learning (Prasetya et al., 2021; Pinandito et al., 2021). The kit in recomposition activities encourages students to build original concept maps that have a well-organized basic structure.

4.2 Analysis of the Quality of ER

Table 2 depicts representative statistics of the quality of ER scores for both groups. The attainment of ER quality scores on control and experimental groups has a similar pattern to the number of ERs. Participants in the experimental group surpassed control groups in terms of minimum, maximum, and median values.

Table 2. *Descriptive Statistics of the Quality of ER Scores for both Groups*

Group	N	Min	Max	Median	Mean	SD
Control group	27	2	6	3.00	3.74	1.35
Experimental group	28	5	27	12.50	14.43	7.54

Mann-Whitney U test was undertaken to examine the significant discrepancy between the groups' quality of ER scores. The statistical analysis results indicated significant differences (Z = -6.160; p = .000 < .05) between the experimental and control group, with Pearson's r of -.831, explaining a large effect size. The mean rank of the control group was 3.741, while that of the experimental group was higher at 14.429.

Students' performance in the experimental group in terms of quality of ER scores consistently outperformed those in control groups. The EKB approach in the experimental group facilitates students to discover ER extensively and attained higher quality semantically. In line with Raud's opinion (2016), the EKB approach encourages learners to have a better quality of personal knowledge than the ESB method. KB’s kit is an important learning activity that provides an essential and solid knowledge structure (Prasetya et al., 2021; Pinandito et al., 2021). Referring to the provided kit, students could understand the teacher's understanding well.

5. Conclusion and Future Work

The highly interconnected concept maps represent proficiency and depth of individual knowledge in the particular domain. The present study sought to reveal the extent of the tightness of the connection between the original map and the additional map on the ESB and the EKB concept mapping tools using ER scores. The experimental results showed that the EKB approach facilitates students to define ER numbers more than ESB. In addition, EKB utilizes KB recomposition which encourages learners to understand the teacher's understanding and produces a solid knowledge structure. In the context of the quality of ER scores, students' achievement using EKB also outperformed those using ESB. However, this is a preliminary study, so further studies are needed to determine the measurements' reliability.

The present study had several limitations that should be considered for further work. First, the number of participants involved in this experiment was relatively small. Thus, future works should consider a larger group of participants to examine the effects of extension concept mapping tools on a broader scale. Second, this study was only conducted at a one-course meeting. However, for more optimal results, continuous experimentation is required. In addition, the experimental groups used the incomplete structure provided, while the control groups did not. For future research, it may be necessary to design a genuinely equitable treatment for both groups. Last, ER achievements may be further analyzed with comprehension tests to obtain useful information.

Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number 19H04227. The first author would like to acknowledge the Islamic Development Bank (IsDB) for a PhD scholarship in partnership with Universitas Negeri Malang (UM), Indonesia.

References

Foley, D., Charron, F., & Plante, J. S. (2018). Potential of the Cogex Software Platform to Replace Logbooks in Capstone Design Projects. *Advances in Engineering Education, 6*(3), n3.

Hirashima, T., Yamasaki, K., Fukuda, H., & Funaoi, H. (2015). Framework of kit-build concept map for automatic diagnosis and its preliminary use. *Research and Practice in Technology Enhanced Learning, 10*(1), 17.

Hirashima, T. (2019). Reconstructional concept map: automatic Assessment and reciprocal reconstruction. *International Journal of Innovation, Creativity and Change, 5*, 669-682.

Jaafarpour, M., Aazami, S., & Mozafari, M. (2016). Does concept mapping enhance learning outcome of nursing students?. *Nurse education today, 36*, 129-132.

Novak, J. D. (2010). Learning, creating, and using knowledge: concept maps as facilitative tools in schools and corporations (2nd ed.)". New York: Routledge. 2010

Pinandito, A., Prasetya, D. D., Hayashi, Y., & Hirashima, T. (2021). Design and development of semi-automatic concept map authoring support tool. *Research and Practice in Technology Enhanced Learning, 16*(1), 1-19.

Prasetya, D. D., Hirashima, T., & Hayashi, Y. (2019). KB-Mixed: A Reconstruction and Improvable Concept Map to Enhance Meaningful Learning and Knowledge Structure. Proceedings of the 26th International Conference on Computers in Education (ICCE 2019). December, 809-812.

Prasetya, D. D., Hirashima, T., Hayashi, Y. (2020). Study on Extended Scratch-Build Concept Map to Enhance Students' Understanding and Promote Quality of Knowledge Structure. *The International Journal of Advanced Computer Science and Applications, 11* (4), 144-153.

Prasetya, D. D., Hirashima, T., Hayashi, Y. (2021). Comparing Two Extended Concept Mapping Approaches to Investigate the Distribution of Students' Achievements. *IEICE TRANSACTIONS on Information and System, 104*(2), 337-340.

Raud, Z., Vodovozov, V., & Lehtla, T. (2016). Teaching, learning, and assessment integration in electronics on the concept map basis. In Innovating with Concept Mapping Proceedings of the Seventh International Conference on Concept Mapping, Tallinn, Estonia (pp. 199-207).

Ruiz-Primo, M. A., Schultz, S. E., & Shavelson, R. J. (1997). On the validity of concept map-base assessment interpretations: An experiment testing the assumption of hierarchical concept maps in science. CRESST.

Ruiz-Primo, M. A., Schultz, S. E., Li, M., & Shavelson, R. J. (2001). Comparison of the reliability and validity of scores from two concept-mapping techniques. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching, 38*(2), 260-278.

Schwendimann, B. A., & Linn, M. C. (2016). Comparing two forms of concept map critique activities to facilitate knowledge integration processes in evolution education. *Journal of Research in Science Teaching, 53*(1), 70-94.

Stoddart, T., Abrams, R., Gasper, E., & Canaday, D. (2000). Concept maps as assessment in science inquiry learning-a report of methodology. *International Journal of Science Education, 22*(12), 1221-1246.

Taricani, E. M., & Clariana, R. B. (2006). A technique for automatically scoring open-ended concept maps. *Educational Technology Research and Development, 54*(1), 65-82.

Turns, J., Atman, C. J., & Adams, R. (2000). Concept maps for engineering education: A cognitively motivated tool supporting varied assessment functions. *IEEE Transactions on Education, 43*(2), 164-173.

Vanides, J., Yin, Y., Tomita, M., & Ruiz-Primo, M. A. (2005). Concept maps. Science Scope, 28(8), 27-31.

Yamasaki, K., Fukuda, H., Hirashima, T., & Funaoi, H. (2010). Kit-build concept map and its preliminary evaluation. Proceedings of ICCE2010, 290-294.

An AES System to Assist Teachers in Grading Language Proficiency and Domain Accuracy Using LSTM Networks

Aditya SAHANI[a*], Forum PATEL[a], Shivani MEHTA[a], Dr Rekha RAMESH[a] & Dr Ramkumar RAJENDRAN[b]

[a]Computer Engineering Department, University of Mumbai, India
[b]Education Technology Department, IIT Bombay, India
**aditya.sahani@sakec.ac.in*

Abstract: Automated Essay Scoring (AES) is a task of automatically grading the students' answers to subjective or essay type questions. AES is an area where assessing the answers rationally is very important. Assessing these subjective answers has always been a challenging process concerning reliability and effort. In such times, where the entire education system has shifted to being online, it becomes necessary to develop a system that assesses students based on their subjective answers. However, the existing AES system primarily focuses on assessing essays on a single dimension that is either grading domain accuracy or grading the language correctness of the answers. Moreover, there are few AES systems to grade student’s responses in the computer science domain. To address these gaps, we propose an AES system to grade the subjective answers of students from the computer science domain. The proposed system grades the student’s responses in two dimensions, namely domain accuracy, and language proficiency. In order to test the system, we collected data from 200 students and manually labeled them for domain accuracy and language proficiency. The system graded the student’s responses automatically with domain accuracy of 89.47 percent and language proficiency of 84.79 percent.

Keywords: Parallel networks, domain accuracy, language proficiency, LSTM, Computer Science, Word Embedding.

1. Introduction

Assessment has been an integral part of the teaching and learning process. Essay type questions are always part of assessments (Dong, Zhang, & Yang, 2017) and such questions offer students an opportunity to demonstrate knowledge, skills, and abilities in a variety of ways such as writing skills and formulating arguments supported with reasoning and evidence (Valenti, Neri, & Cucchiarelli, 2003). However, because of the large number of student participation in assessments, manual evaluation and grading of answers to these essay type questions is a challenging task for the teachers. Manual evaluation by multiple teachers may also introduce inconsistent or erroneous grading because of mutual disagreements (Dasgupta, Naskar, Dey, & Saha, 2018).

To address this challenge automated essay scoring (AES) has been explored for over 50 years as a part of natural language processing. The existing works in AES were primarily developed to assess the essays on a single dimension such as either assessing domain knowledge or language proficiency. In order to provide detailed feedback to students and measure the outcome of all the objectives of subjective assessments, it is necessary to grade them on both domain accuracy and language proficiency. There exist few systems to detect both domain accuracy and language proficiency. However, they provide the final score as a combination/average of the domain and language accuracy. To address this gap, we propose a system that grades essays on mainly two parameters: domain accuracy and language proficiency.

The related systems with AES implement word embedding, transfer learning and feature engineering to score the answers (Hussein, Hassan, & Nassef, 2019) (Hussein, A., & Nassef, 2020). Many systems with respect to AES have been using LSTMs and CNNs to find meaning behind text and score it (Taghipour & Ng, 2016)(Hussein, A., & Nassef, 2020) (Cao, Jin, Wan, & Yu, 2020). A Siamese Bidirectional LSTM based Regression was designed for grading the answers of the Computer Science domain (Prabhudesai, Arya, Duong, 2019).

A few systems work in the CS domain (Ndukwe, Amadi, Nkomo, & Daniel, 2020) (Hussein, Hassan, & Nassef, 2019) and hence, there is limited availability of dataset from CS domain. Very few systems predict a score on English language proficiency (Ndukwe, Amadi, Nkomo, & Daniel, 2020) (Zhao, Zhang, Xiong, Botelho, & Heffernan, 2017). Moreover, most of the papers are improving the performance over the existing ASAP dataset.

Our proposed system uses a parallel Long Short Term Memory (LSTM) network, one for checking the domain accuracy of the answer and another for checking the language proficiency. We selected LSTM networks because they are widely used in recent related works and it utilizes the long-distance dependencies in the answers (Taghipour & Ng, 2016). From the existing AES systems, we found that there is no dataset in the CS domain that can be used for scoring. To address scarcity of dataset in CS domain, we created a dataset. The dataset is called the UM dataset and has answers Data Structures (DS). 200 students were given four essay-type questions as part of their assignments. Students' answers were graded by the instructors for domain and language accuracy and were given a score using a rubric designed by subject experts.

To train and test our system, a corpus of words was developed related to the Data structures domain using three standard textbooks from the university curriculum. The words in the corpus are then converted into word embedding using the Gensim Library, the embedding size is set to 300 dimensions The proposed AES system graded the student's response with a domain accuracy of 89.47 percent, language proficiency of 84.7 percent.

Since the UM dataset is small, to test the performance of our system on a large dataset we searched for an existing dataset in the DS course. We found a database from the University of North Texas (UT dataset). The dataset had short answers from the Data Structures domain. However, the UT dataset was associated with only domain scores and the rubrics used for grading were unavailable. Hence, we graded the UT dataset for language proficiency by experts using the rubrics that we developed. The proposed system graded the answers in UT dataset with a domain accuracy of 89.43 percent and language proficiency of 89.92 percent. The results indicate that our proposed dual LSTM network can perform well on both small and large datasets.

The rest of the paper is divided as follows. Section 2 discusses our proposed system. Section 3 describes the detailed implementation of our system and the results of the system are provided in section 4. Following this, section 5 throws light on the discussion and conclusion of the system.

2. Proposed Methodology

We use a parallel LSTM networks for automated grading of students' subjective answers to predict a separate individual score for domain accuracy as well as language proficiency for these answers. Figure 1 shows the block diagram of the proposed system. This trained word2vec model is stored and then fed to the LSTM model where the domain accuracy and language proficiency score is calculated.

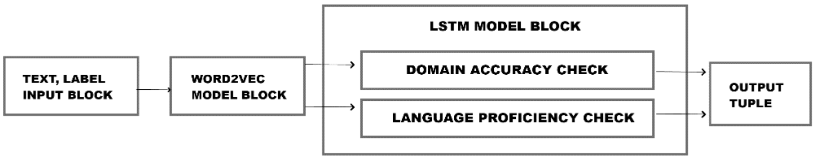


Figure 1. System Flow Diagram.

3. Implementation

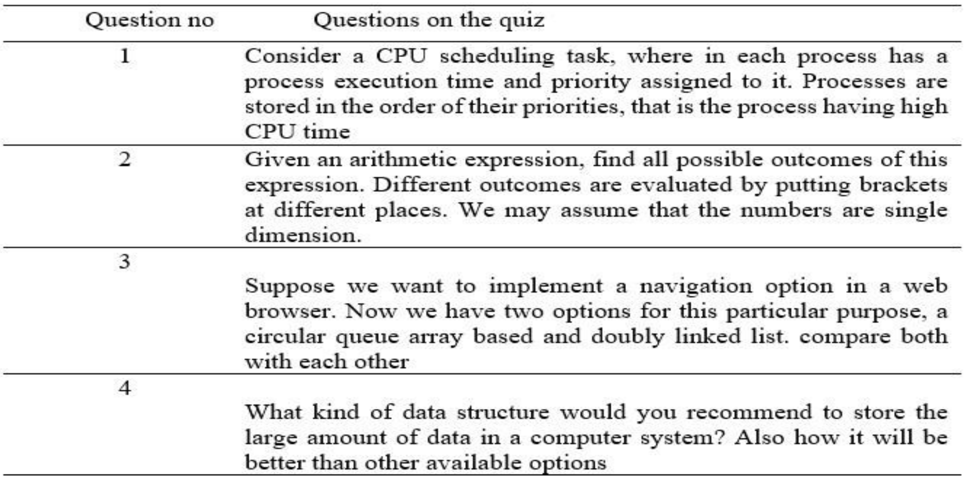
The model takes in the answers of the data structures course graded on the domain accuracy and language proficiency as its input. The input fed is the word embeddings. The system uses a parallel LSTM network to score the answers. After successfully training and testing the model, the scores are displayed in the form of a tuple. The functions of each block are described in the following subsections.

The input block of the system takes in the answers of the data structures course that are graded by experts on both domain accuracy and language proficiency. The block performs the data cleaning by removing the nulls and performs exploratory data analysis on the subjective answers.. The dataset and pre-processing are explained in detail below.

We created our data set called the UM (University of Mumbai) dataset, consisting of descriptive answers for the data structures course offered in the second-year engineering curriculum. Students were given a set of four questions to solve in one hour as part of their assignment. The questions are shown in Table 1. 200 students participated in the assignment. These answers were collected online.

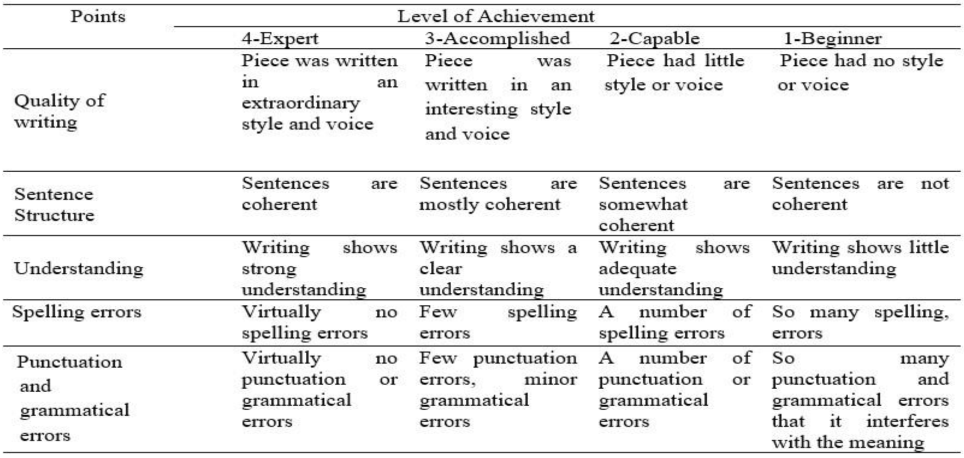
Two domain experts who taught the course three times graded the answers on domain accuracy. To have consistency in grading, a rubric was designed as shown in Table 3 and the inter-rater reliability achieved with Cohen kappa was 0.82. Table 2 talks about the language rubric that has been used to score the answers on language proficiency by an Expert.

Table 1. *Assignment questions for creating UM dataset*



Question no	Questions on the quiz
1	Consider a CPU scheduling task, where in each process has a process execution time and priority assigned to it. Processes are stored in the order of their priorities, that is the process having high CPU time
2	Given an arithmetic expression, find all possible outcomes of this expression. Different outcomes are evaluated by putting brackets at different places. We may assume that the numbers are single dimension.
3	Suppose we want to implement a navigation option in a web browser. Now we have two options for this particular purpose, a circular queue array based and doubly linked list. compare both with each other
4	What kind of data structure would you recommend to store the large amount of data in a computer system? Also how it will be better than other available options

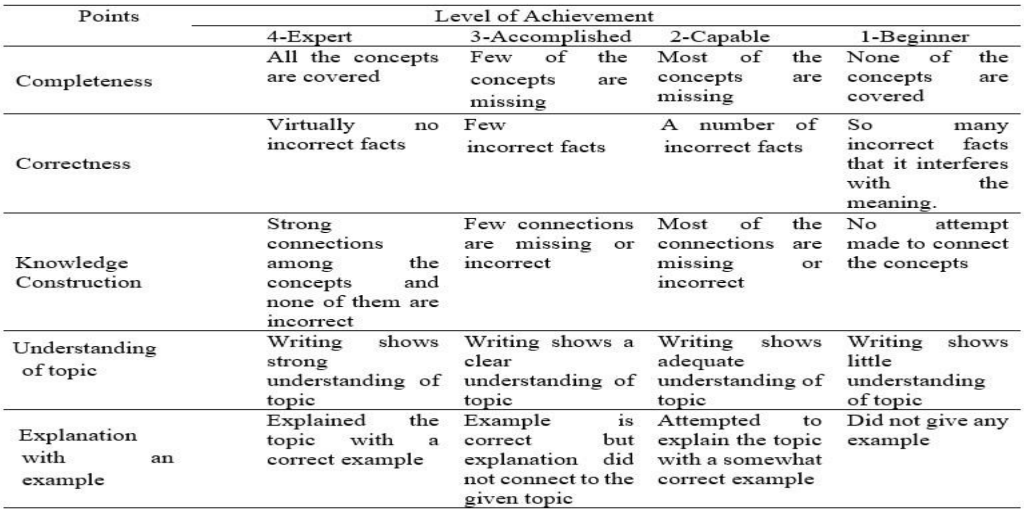
Table 2. *Rubrics for accessing Language Proficiency*



Points	Level of Achievement			
	4-Expert	3-Accomplished	2-Capable	1-Beginner
Quality of writing	Piece was written in an extraordinary style and voice	Piece was written in an interesting style and voice	Piece had little style or voice	Piece had no style or voice
Sentence Structure	Sentences are coherent	Sentences are mostly coherent	Sentences are somewhat coherent	Sentences are not coherent
Understanding	Writing shows strong understanding	Writing shows a clear understanding	Writing shows adequate understanding	Writing shows little understanding
Spelling errors	Virtually no spelling errors	Few spelling errors	A number of spelling errors	So many spelling, errors
Punctuation and grammatical errors	Virtually no punctuation or grammatical errors	Few punctuation errors, minor grammatical errors	A number of punctuation or grammatical errors	So many punctuation and grammatical errors that it interferes with the meaning

We created a corpus that consisted of words from the domain from three standard books from the university curriculum. We use the gensim library to create the word2vec model. The model is trained with an embedding size of 300 and uses the skip-gram model (Agarap, 2019). After this trained model is stored, when we input the answers the model checks for new words and adds them to the word2vec model corpus.

Table 3. *Rubrics for accessing Domain Accuracy*



Points	Level of Achievement			
	4-Expert	3-Accomplished	2-Capable	1-Beginner
Completeness	All the concepts are covered	Few of the concepts are missing	Most of the concepts are missing	None of the concepts are covered
Correctness	Virtually no incorrect facts	Few incorrect facts	A number of incorrect facts	So many incorrect facts that it interferes with the meaning.
Knowledge Construction	Strong connections among the concepts and none of them are incorrect	Few connections are missing or incorrect	Most of the connections are missing or incorrect	No attempt made to connect the concepts
Understanding of topic	Writing shows strong understanding of topic	Writing shows a clear understanding of topic	Writing shows adequate understanding of topic	Writing shows little understanding of topic
Explanation with an example	Explained the topic with a correct example	Example is correct but explanation did not connect to the given topic	Attempted to explain the topic with a somewhat correct example	Did not give any example

The LSTM model consists of five layers. The recurrent dropout throughout the model is set to 40 percent and dropout is set to 50 percent. The first layer is the embedding layer, second LSTM layer consists of 300 neurons, input shape is set to 1,300 and the return sequence parameter is set to true. The third layer is an LSTM layer with 64 neurons, fourth layer is a dropout layer and last is a dense layer to take output.

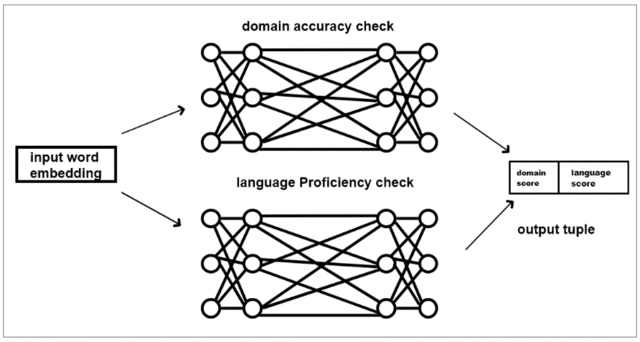


Figure 2. Parallel LSTM Network Model.

The activation function used is a rectified linear (ReLU) unit throughout the model (Agarap, 2019). We used mean square error, mean absolute error and cohens kappa metrics to evaluate the model. The input is optimized using the Adam optimizer (Kingma & Ba, 2017). The system uses a parallel LSTM network that work to score the answers in Figure 2.

The output block of the system consists of the results of the two parallel LSTM networks into a single tuple {Score 1, Score 2}. The first value in the tuple will be the score of domain accuracy and the second will be the language proficiency.

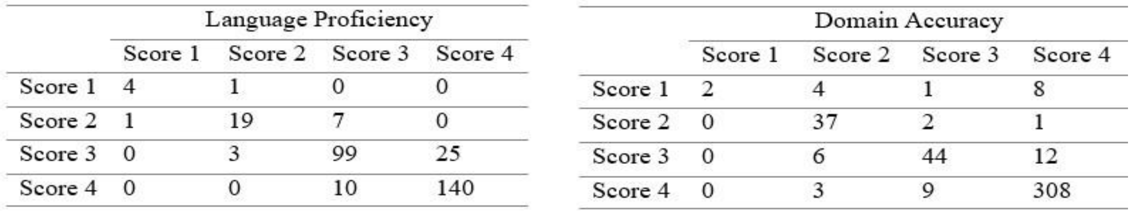
4. Results

The parallel LSTM model is first tested on the UM dataset. This dataset is fed to the model for both domain accuracy and language proficiency. The model is trained for 10 folds of cross-validation and

50 epochs each fold. Table 4 indicates the domain accuracy confusion matrix when the model is tested on the UM dataset and table 5 indicates the language proficiency. After successfully training the model, its performance is summarized as shown in Table 4.

After evaluation of the model, it produces a mean square error (mse) of 0.4 for domain accuracy and 0.2 for language proficiency. A quadratic weighted kappa (qwk) of 76.5 and 86.2 for domain accuracy and language proficiency, a precision score of 68.2 and 80, a recall score of 86.7 and 83, the overall accuracy of the model is 89.47 for domain accuracy and 84.7 for language proficiency.

Table 4. *The Confusion matrix for the language proficiency & domain accuracy of the dataset UM*

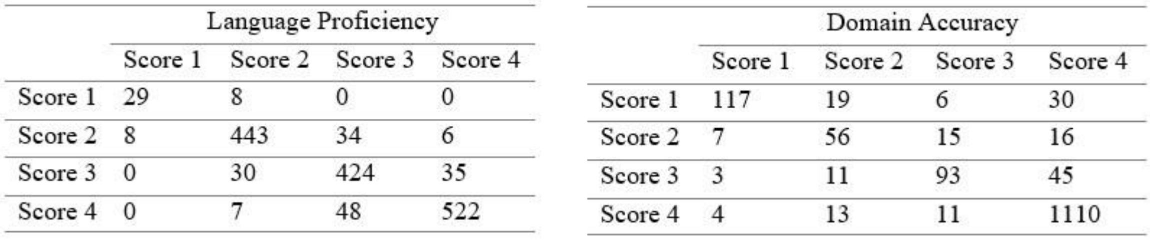


	Language Proficiency			
	Score 1	Score 2	Score 3	Score 4
Score 1	4	1	0	0
Score 2	1	19	7	0
Score 3	0	3	99	25
Score 4	0	0	10	140

	Domain Accuracy			
	Score 1	Score 2	Score 3	Score 4
Score 1	2	4	1	8
Score 2	0	37	2	1
Score 3	0	6	44	12
Score 4	0	3	9	308

The UM dataset is smaller in size, due to which we also tested it on the UT dataset to check the generalizability of our model. The UT dataset created by the University of North Texas consists of questions from the Data Structures course given as part of a course assignment (Mohler, Bunescu, & Mihalcea, 2011) (Mohler & Mihalcea, 2009). The dataset in all had 1400 answers rated on domain accuracy in the range of 1(Beginner) to 4(Advanced). The UT dataset does not have a language score assigned to the answers. Hence, we had an expert (doctorate in English literature) who graded the answers on language proficiency using a rubric shown in Table 2.The model runs for 10 folds of cross-validation with 50 epochs each fold. Table 5 indicates the domain accuracy confusion matrix when the model is tested on the UT dataset and table 8 indicates the language proficiency confusion matrix.

Table 5. *The Confusion matrix for the language proficiency & domain accuracy of the UT dataset*



	Language Proficiency			
	Score 1	Score 2	Score 3	Score 4
Score 1	29	8	0	0
Score 2	8	443	34	6
Score 3	0	30	424	35
Score 4	0	7	48	522

	Domain Accuracy			
	Score 1	Score 2	Score 3	Score 4
Score 1	117	19	6	30
Score 2	7	56	15	16
Score 3	3	11	93	45
Score 4	4	13	11	1110

After evaluation of the model, it produces a mean square error (mse) of 0.1 for domain accuracy and 0.3 for language proficiency. A quadratic weighted kappa (qwk) of 81 and 92 for domain accuracy and language proficiency, a precision score of 70.7 and 80, a recall score of 73.3 and 86.5, the overall accuracy of the model is 89.43 for domain accuracy and 89.92 language proficiency.

5. Discussion and Conclusion

We developed a system that uses parallel LSTM networks to grade students’ subjective answers. Due to the limited availability of data in the CS domain, we created our own UM dataset by conducting an assessment consisting of essay type questions for the second year Engineering students from University of Mumbai in the data structures course. The answers are rated by experts for domain accuracy and language proficiency. The expert raters used two different rubrics to grade these answers and the grades ranged between 1(Beginner) to 4(Expert) as explained previously in section 4.1.1.

The system was also tested on the bigger dataset (UT) which we got from the University of North Texas (Mohler, Bunescu, & Mihalcea, 2011) (Mohler & Mihalcea, 2009) to check how the model would generalize for a large dataset. The results show that it performs well across both datasets for domain accuracy and language proficiency. We plan to share UM data publicly so that other researchers can use and build an AES system around it.

Currently, there is very limited availability of datasets in the CS domain. The UM dataset consists of only subtopics of the data structures course whereas the UT dataset has answers from all the topics of the course. The results of the UM dataset, when evaluated and analyzed, portray that the dataset might be unbalanced. The system performance can be improved if more data is available for training. The limitation of the system is that it can grade the answers only with integer values. Another limitation of the system is that it cannot provide feedback to the user on why the answer has been given a particular score.

In future, this system could be generalized to more domains to include other courses of CS Domain. We can also dive deeper into NLP and try including various text features that can be used with these word vectors to generate a more accurate system. The system could incorporate a self-explanatory feedback mechanism based on the rubrics and help the students in self-learning and improvement.

References

Agarap, A.0 F. (2019, February 07). Deep Learning using Rectified Linear Units (ReLU). https://arxiv.org/abs/1803.08375

Cao, Y., Jin, H., Wan, X., & Yu, Z. (2020). Domain-Adaptive Neural Automated Essay Scoring. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. doi:10.1145/3397271.3401037

Dasgupta, T., Naskar, A., Dey, L., & Saha, R. (2018). Augmenting Textual Qualitative Features in Deep Convolution Recurrent Neural Network for Automatic Essay Scoring. Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications. doi:10.18653/v1/w18-3713

Dong, F., Zhang, Y., & Yang, J. (2017). Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring. Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). doi:10.18653/v1/k17-1017

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. PeerJ Computer Science, 5. doi:10.7717/peerj-cs.208

Hussein, M. A., A., H., & Nassef, M. (2020). A Trait-based Deep Learning Automated Essay Scoring System with Adaptive Feedback. International Journal of Advanced Computer Science and Applications, 11(5). doi:10.14569/ijacsa.2020.0110538

Kingma, D. P., & Ba, J. (2017, January 30). Adam: A Method for Stochastic Optimization. Retrieved from https://arxiv.org/abs/1412.6980

Mohler, M., & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on - EACL 09. doi:10.3115/1609067.1609130

Mohler, M., Bunescu, R., & Mihalcea, R. (2011). Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. Retrieved from https://aclanthology.org/P11-1076/

Ndukwe, I. G., Amadi, C. E., Nkomo, L. M., & Daniel, B. K. (2020). Automatic Grading System Using Sentence-BERT Network. Lecture Notes in Computer Science Artificial Intelligence in Education, 224-227. doi:10.1007/978-3-030-52240-7_41

Prabhudesai, A., & Duong, T. N. (2019). Automatic Short Answer Grading using Siamese Bidirectional LSTM Based Regression. 2019 IEEE International Conference on Engineering, Technology and Education (TALE). doi:10.1109/tale48000.2019.9226026

Taghipour, K., & Ng, H. T. (2016). A Neural Approach to Automated Essay Scoring. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. doi:10.18653/v1/d16-1193

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An Overview of Current Research on Automated Essay Grading. Journal of Information Technology Education: Research, 2, 319-330. doi:10.28945/331

Zhao, S., Zhang, Y., Xiong, X., Botelho, A., & Heffernan, N. (2017). A Memory-Augmented Neural Model for Automated Grading. Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale. doi:10.1145/3051457.3053982

Conceptual Level Comprehension Support of the Object-Oriented Programming Source-Code Using Kit-Build Concept Map

Nawras KHUDHUR[a*], Pedro Gabriel Fonteles FURTADO[a], Aryo PINANDITO[a], Shimpei MATSUMOTO[b], Yusuke HAYASHI[a] & Tsukasa HIRASHIMA[a]

[a]*Graduate School of Advanced Science and Engineering, Hiroshima University, Japan*
[b]*Faculty of Applied Information Science, Hiroshima Institute of Technology, Japan*
[*]nawras@lel.hiroshima-u.ac.jp

Abstract: Object-oriented programming (OOP) is a modern model of programming languages and an important module for many programming courses in academics. Not only do educators have trouble teaching OOP concepts but students are also reported to having trouble comprehending those concepts. The difficulty lies in dealing with abstract concepts and finding a relationship between the textbook explanations and the application of these concepts. Several works try to approach this problem, but they lack connecting the OOP concepts with its implementation in the source-code. In this research, we propose a new visualization form using concept maps to combine the OOP concepts with its' source-code to promote OOP concept comprehension. The proposed visualization is called the conceptual representation of the source-code (CRS). CRS unites the source-code statements and the OOP concepts into one comprehensible diagram. A concept map recomposition activity with Kit-Build is used to implement the CRS. We have conducted an experiment on university students to verify the learning effects and use of the proposed method. The results show a significant improvement in immediate learning by comparing before/after activity test-scores. In addition, students showed a positive impression and intention about using the tool during their studying of OOP by answering a questionnaire. The research findings shed light on a promising aspect of teaching OOP concepts in programming courses.

Keywords: Concept comprehension, conceptual level representation, concept map, object oriented programming, OOP concepts, kit build, concept visualizer.

1. Introduction

Computer programming is the main subject of study curriculum in computer-science related fields and even some high-schools that provide programming classes. Being a major part of computer programming, object-oriented programming (OOP) is a recent paradigm of programming languages. In OOP, programs consist of classes and objects. This structure is beneficial to divide problems into smaller pieces thus making the problem-solving more natural and the code reusable. OOP consists of several strongly interrelated concepts. Armstrong (2006) lists the concepts as object, class, method, message passing, inheritance, polymorphism, encapsulation, abstraction, instantiation, and modeling. Teaching these concepts and its comprehension in terms of actual coding is shown to be a difficult task for both educators and students.

The difficulty of OOP concept comprehension is identified in the literature (Kaczmarczyk et al., 2010; Sorva, 2018). Relational problems are also described where a study shows students find it difficult to comprehend the relationship among different concepts (Sajaniemi et al., 2008). Lack of active practice and suitable teaching tools are one of the reasons why it is hard to teach students about OOP concepts (Sarpong et al., 2013).

One familiar way to represent some of the OOP concepts is to use UML class diagrams. But in UML, the general explanations of the OOP concepts and how it is related to the source-code is not represented. Students are left with a code structure which is helpful but not enough to comprehend OOP concepts especially when students are less experienced with UML class diagrams (Gravino et al., 2015). In some courses, program visualizers (PV) are implemented as a tool to support OOP concept comprehension. PVs are tools that show the run-time behavior of a program when executed. Despite its usefulness, PVs could not fill the gap of OOP concept comprehension, mainly because of the low engagement structure of the PV and its limitation in visualizing OOP concepts and its' relationships (Sorva et al., 2013). Thus, educators need a tool to create conceptual level activities that can correlate the general OOP concept explanations with the actual source-code implementation effectively.

To approach this goal, we investigate a way to combine the OOP concepts and the source-code into one diagram using concept map (CMAP) (Novak, 2005). We call this combination representation *Conceptual Representation of The Source-code (CRS).* Creating such a relational view between OOP concepts and its' actual use in source-code is not proposed before to the extent of our knowledge. By this combination, we aim to expose the learner to a productive view of the theory and practice of OOP concepts and encourage learners to interact with it actively. One issue with conventional CMAP is that each learner tends to construct the map for the targeted knowledge differently since the map depends on learners' conceptual understanding which can vary from learner to learner. One way to transform concept mapping into a more manageable and controllable yet effective activity for educators is using Kit-Build (KB) recomposition (Hirashima et al., 2015). KB is one type of CMAP that focuses on expert map recomposition instead of free map creation. In KB the learners are provided with a kit of concepts and links. Learners' goal is to recompose the CMAP in the same way the expert built it. This activity is called *concept map recomposition*.

In this paper, we investigate the possibility of using KB concept map to create an activity that unifies OOP source-code with its concepts i.e., implementing CRS. In addition, we aim to consider its impact on OOP concept comprehension.

2. Kit-Build Concept Map Recomposition

Concept map was first introduced by Novak (Novak, 2005) to evaluate students' conceptual learning and progress. In CMAP, concepts are expressed as nodes. These concepts are then connected to each other using labels to form a meaningful proposition. It is confirmed by many studies that it can promote the learning process in a variety of subjects and specialties (Wang and Chen, 2018; Balim, 2013).

CMAP comes in various forms such as scratch map (SM) and closed concept map (CCM) (Furtado et al., 2019). SM is a traditional concept mapping where learners start from an empty layout and build the concept map gradually. SM has unconstrained variability as it reflects each individual. In contrast, CCM provides a limited map building environment where learners are given a selected set of concepts and labels to choose from while building the map. However, the learners are free to make any proposition that they assume is valid using the provided set.

A more restricted type of CCM introduced by Hirashima et al. (2015) called Kit-Build (KB). KB asks learners to recompose the concept map instead of building it. By recompose, it means to re-connect a concept map from a kit of concepts and links of a pre-built concept map (expert map). The steps of a simple KB activity are as follows: 1) The expert creates a concept map for a material. 2) The expert concept map is then decomposed to its basic parts by removing the connections, thus creating a kit of concepts and links. 3) The kit is given to the learners to recompose it to the expert map. In KB, it is possible to have an exact map comparison between learners' map and the expert map since the same map pieces are used to make learner map and expert map is used as a reference. This comparison allows instructors to pinpoint the difficult parts of the lecture and give more accurate feedback to the learners (Sugihara et al., 2012). The validity and reliability of the KB diagnosis tool compared to the traditional map evaluation have been verified by past research (Wunnasri et al., 2018, 2017). KB also makes it possible to give automatic feedback to the learners while recomposing the map such as highlighting different propositions compared to the expert map. Another study showed that using expert map recomposition let the learners get a broader and deeper knowledge comprehension compared to scratch map building (Prasetya et al., 2021).

Despite of these many studies about concept maps and particularly KB, there are no investigations about using it in technical comprehension tasks such as to represent source-code and its concepts up to the authors' knowledge.

In this research, we used KB to implement CRS. The screenshot of the KB is shown in Figure 1. In KB, labels have two connectors colored red and blue, which appear only when the label is selected. The red connector refers to the source of the relationship, while the blue connector means the target of the relationship. Porpositions can be made by connecting these connectors to the concepts. Labels can make one-to-many relationships with concepts. Hence, the number of connectible targets is shown inside the blue connectors circle.

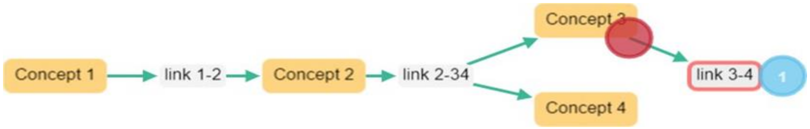


Figure 1. Kit-Build Example.

3. CRS Concept Map

The structure of an object-oriented (OO) source-code can be divided into two sections, internal and external structure. The internal structure refers to the statements of the source-code such as method/variable definitions and so on. It is explicitly visible to the learner i.e. the learner can just read through the source-code. Another feature of the internal structure is that it can go differently compared to another OO source-code, since the structure is the written text itself. In contrast, the external structure describes the OOP concepts such as inheritance, polymorphism, etc. These concepts are not directly visible in the OO source-code but its' implementation is realized in it. Moreover, the "**fact"**s of these OOP concepts are not dependent on the written text itself. In this sense, two different OO source-code can implement these OOP concepts in the same manner.

The objective of CRS concept map is to visualize both structures in one diagram and act as an intermediary between the two structures. Bridging these two structures allows the learner to connect the concepts of OOP to its' actual implementation in the source-code. Consequently, it can promote the conceptual interrelationships and how they affect each other. Another use of CRS is to use it as an evaluation map to measure the quality of an OO source-code by representing a given source-code in CRS and focus on what OOP concepts cannot be represented. This concludes that learners code does not implement these OOP concepts.

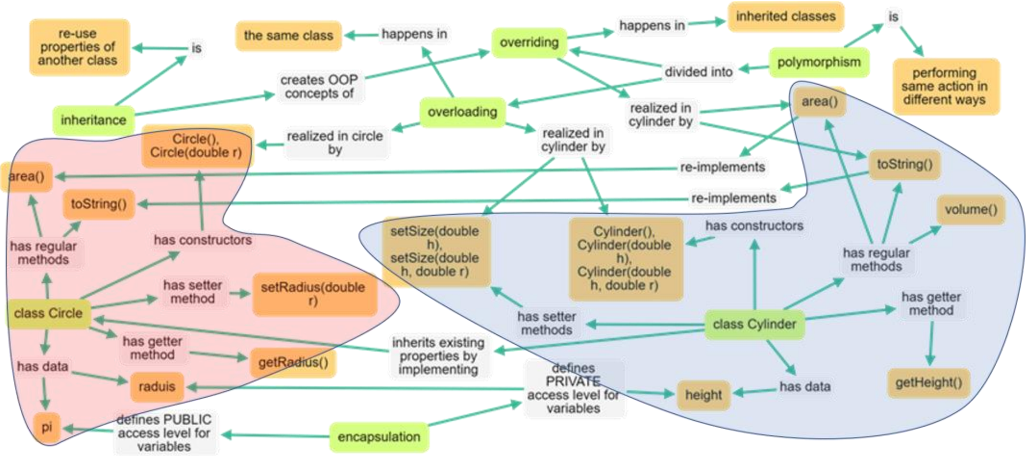


Figure 2. The Expert Goal-map for Learning OOP Concepts in a Source-code.

Figure 2 shows the proposed implementation of CRS concept map. The concept map is based on a source-code that implements a set of the OOP concepts[1]. The source-code consists of two classes Circle and Cylinder. Both classes contain multiple constructors. The Cylinder class inherits the Circle class and overrides two methods of Circle, namely area and toString. Concepts included in the source-code are inheritance, polymorphism, encapsulation, class, and composition. The red and blue

[1] https://git.io/JtKt6

regions in Figure 2 represents the internal structure of the class Circle and Cylinder respectively. These basic elements are the parts of the source-code that can be noticed easily by any learner, but it is not self-explainable. Several propositions are used to create the internal structure. These propositions act like annotations to support the fundamental source-code comprehension.

After grasping the fundamentals, the next target would be bridging it to the external structure of the source-code. The propositions outside the colored regions are the key propositions in CRS that act as a bridge to connect the major sections of the source-code to the OOP concepts. Hence, it enables the learner to foster the OOP concepts in the practical environment. The bridging propositions indicate what parts in the source-code represent the corresponding OOP concept. Moreover, CRS wraps up and provides the big picture of the implemented OOP concepts enticing the learner to track the interrelationship among different OOP concepts. To give an example, the CRS can tell why “overriding” can occur when inheritance is implemented, but it is not possible with a standalone class.

The CRS can be extended to target different goals. For instance, an “object” of type Cylinder can be added to CRS to reveal the access restrictions of an object toward different class variables and methods. It can also be modified to meet the understanding level of the students by adding more details of the OOP concepts when it is the first time to introduce OOP concepts.

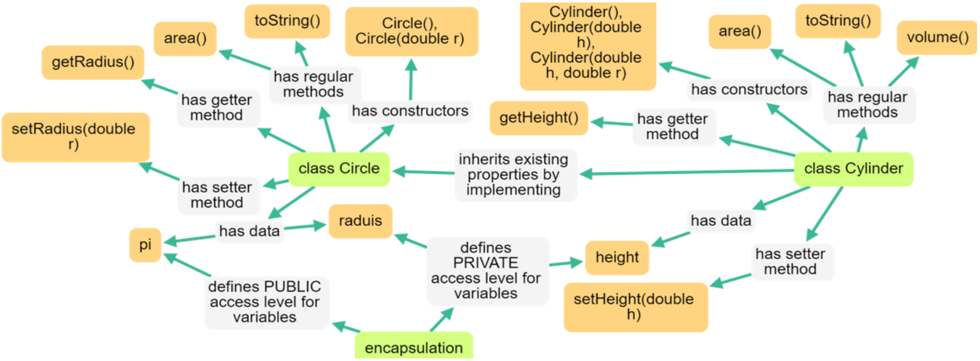


Figure 3. The Expert Goal-map Adopted in The Experiment.

4. Research Methodology

To investigate the effectiveness of CRS, a quasi-experimental design was utilized. Students performed a pre-test, recomposed the map, performed the post-test. At the end of the activity,students were asked to fill in a questionnaire about the activity.

4.1 Participants

The participants were 49 undergraduate third-year university students, majored in computer science. The experiment was conducted during their regular class and KB is used as a part of class-teaching material. Thus, we could not prepare a control group. Students were free to discontinue the experiment at any stage. Particularly, out of 49 students, 31 students completed the experiment and only the data for those 31 students were included in the analysis.

4.2 Materials

For this experiment, two materials were prepared. The first material was online lecture notes OOP concepts. The second material was a source-code explained in Section 3. An expert map implementing CRS was prepared which was about the realized OOP concepts in the given source-code. However, the map was simplified for this experiment to include fewer concepts and details, since the class time was very limited. The used expert map in the experiment is shown in Figure 3.

4.3 Procedure

The class instructor started the experiment by explaining the KB to the students and let them to build the training map to get familiar with the tool and its features. After that, students tested for their basic knowledge about the concepts of OOP given a source-code as a reference by answering multiple-choice questions about the concepts that were applied in the given source-code. This pre-test session lasted for 5 minutes. Afterward, material about the concepts of OOP given to the students to read and briefly explained by the instructor in 10 minutes. Then students were asked to recompose the kit using KB in 25 minutes. During concept map recomposition, students were allowed to look at the source-code. The KB tool had feedback feature to evaluate learners' map. The feedback reports the wrong propositions made by the learner that does not exist in the expert map. Students did a post-test afterward.

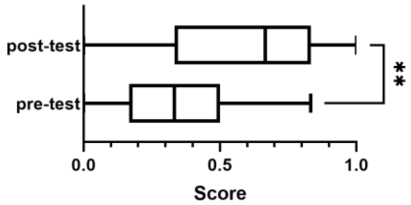


Figure 4. Comparison of Pre-test Score with Post-test Score.

4.4 Learning Effect of CRS Recomposition on OOP Concept Comprehension

We run the Fligner-Killeen test of homogeneity of variances to make sure there is no selection bias for students who decided to finish the experiment successfully. The result shows that the students were homogeneous with a p-value of 0.09639. To measure the learning outcome of using the CRS, we have compared the post-test scores against pre-test scores. The scores failed the Shapiro-Wilk normality test due to small number of participants. Thus, we used the Wilcoxon matched-pairs signed-rank test to measure the difference. Figure 4 shows the comparison results. The medians of pre-test score and post-test score were 0.33 and 0.66, respectively. A Wilcoxon Signed-rank test showed a significant difference between the post-test and pre-test scores (W=229, Z=-2.6754, p-value=0.006, r =0.346). The result suggests that using CRS recomposition can promote learning OOP concepts adequately.

4.5 Students' Feedback on Applying CRS in The Class

A questionnaire was given consisted of six 6-Likert scale questions to measure the students' likeness and expectation regarding the use of the CRS in learning OOP. Students showed a positive impression in using CRS, showing the average score for likeness and expectation 4.2 and 4.4 respectively. We can estimate that this activity was a little odd for students since it was different from their usual learning methods but still useful. The students' expectation for the CRS recomposition is high and positive. The students mostly agree that this activity will help them in learning OOP concepts effectively.

5. Concluding Remarks

In this study, we have presented a novel way to visualize OOP concepts and combine it to the OOP-based source-code. The post-test score results showed a significant improvement in students' OOP concept comprehension after the recomposition activity of the proposed concept map. Additionally, the questionnaire analysis of students' feedback shows that learning with the proposed activity is memorable and friendly as well as fun to some extent.

This study raises a new insight toward the perception of researchers and educators on OOP concept comprehension solutions coupled with source-code. The proposed approach has potential in various teaching aspects. It can be used in creating OOP-based activities to promote OOP comprehension. It also allows educators to create conceptual-based activities when teaching OOP.

One future work is to set a control group to compare it to other similar methods and to generalize the findings in this study. Another essential future work is considering the CRS as a source-code qualifier in terms of OOP concepts. It can be used by the learners to self-evaluate their source-code or by the teacher to group-evaluate the learners' source-code.

Acknowledgments

This research is partially supported by JSPS KAKENHI Grant Numbers 19H04227.

References

Armstrong, D. J. (2006). The quarks of object-oriented development. *Communications of the ACM*, 49(2), 123-128.

Balim, A. G. (2013). Use of technology-assisted techniques of mind mapping and concept mapping in science education: a constructivist study. *Irish Educational Studies*, 32(4), 437-456.

Furtado, P. G. F., Hirashima, T., & Hayashi, Y. (2019). Reducing cognitive load during closed concept map construction and consequences on reading comprehension and retention. *IEEE Transactions on Learning Technologies*, 12(3):402-412.

Gravino, C., Scanniello, G., & Tortora, G. (2015). Source-code comprehension tasks supported by UML design models: Results from a controlled experiment and a differentiated replication. *Journal of Visual Languages and Computing*, 28, 23-38.

Hirashima, T., Yamasaki, K., Fukuda, H., & Funaoi, H. (2015). Framework of kit-build concept map for automatic diagnosis and its preliminary use. *Research and Practice in Technology Enhanced Learning*, 10(1), 17.

Kaczmarczyk, L. C., Petrick, E. R., East, J. P., & Herman, G. L. (2010). Identifying student misconceptions of programming. *SIGCSE'10 - Proceedings of the 41st ACM Technical Symposium on Computer Science Education*, 107-111.

Novak, J. D. (2005). Results and implications of a 12-year longitudinal study of science concept learning. *Research in Science Education*, 35(1), 23-40.

Prasetya, D. D., Hirashima, T., And Hayashi, Y. (2021). Comparing two extended concept mapping approaches to investigate the distribution of students' achievements. *IEICE Transactions on Information and Systems*, E104.D(2):337-340.

Sajaniemi, J., Kuittinen, M., & Tikansalo, T. (2008). A study of the development of students' visualizations of program state during an elementary object-oriented programming course. *ACM Journal on Educational Resources in Computing*, 7(4).

Sarpong, K. A.-M., Arthur, J. K., and Amoako, P. Y. O. (2013). Causes of failure of students in computer programming courses: The teacher-learner perspective. *International Journal of Computer Applications*, 77(12).

Sorva, J. (2018). Misconceptions and the beginner programmer. *Computer Science Education: Perspectives on Teaching and Learning in School* (pp. 171-187). Bloomsbury Publishing.

Sorva, J., Karavirta, V., & Malmi, L. (2013). A review of generic program visualization systems for introductory programming education. *ACM Transactions on Computing Education*, 13(4).

Sugihara, K., Osada, T., Hirashima, T., Funaoi, H., & Nakata, S. (2012). Experimental evaluation of kit-build concept map for science classes in an elementary school. *Proceedings of the 20th International Conference on Computers in Education*, ICCE 2012, 17-24.

Wang, S. P., & Chen, Y. L. (2018). Effects of multimodal learning analytics with concept maps on college students' vocabulary and reading performance. *Educational Technology and Society*, 21(4), 12-25.

Wunnasri, W., Pailai, J., Hayashi, Y., & Hirashima, T. (2017). Reliability investigation of automatic assessment of learner-build concept map with kit-build method by comparing with manual methods. *International Conference on Artificial Intelligence in Education*, 418-429.

Wunnasri, W., Pailai, J., Hayashi, Y., and Hirashima, T. (2018). Validity of kit-build method for assessment of learner-build map by comparing with manual methods. *IEICE Transactions on Information and Systems*, E101.D(4):1141-1150.

Wunnasri, W., Pailai, J., Hayashi, Y., & Hirashima, T. (2018). Validity of kit-build method for assessment of learner-build map by comparing with manual methods. *IEICE Transactions on Information and Systems*, E101D(4), 1141-1150.

Automatic Classification of MOOC Forum Messages to Measure the Quality of Peer Interaction

Urvi SHAH[a*], Richa RAMBHIA[a], Prakruti KOTHARI[a], Rekha RAMESH[a] & GargiBANERJEE[b]
[a]*Department of Computer Engineering, Shah & Anchor Kutchhi Engineering College, University of Mumbai, India*
[b]*Educational Technology Programme, IIT Bombay, India*
*umshah99@gmail.com

Abstract: Discussion forum is an integral part of many MOOCs as it provides a platform for peer interaction among learners. The quality of peer interaction is an indicator of the potential for peer learning. Thus, quality of peer interaction provides instructors with an actionable insight into the extent of critical or higher level thinking that learners are engaged in and is a measure of the learning effectiveness of the course. It is daunting for instructors to manually analyze the forum messages to gain this insight. To address this issue, we attempted to develop a system for automatic classification of forum messages that will inform instructors on the quality of peer interaction happening in the forum. Our system classifies messages into predefined classes based on the Interaction Analysis Model phases. We explored and implemented multiple machine learning models. A general accuracy of 95%-97% was observed among the models and no model outperformed the other by a great margin. The needfor such a system has become all the more relevant in the current Covid-19 pandemic situation, where all physical classrooms have had to migrate to an online setting.

Keywords: Massive open online courses; discussion forum; peer interaction; automatic classification; machine learning; neural network

1. Introduction

Discussion forums (DF) offer a platform for asynchronous communications that facilitates interactions and communications among learners and instructors, and it also helps learners build a community within the MOOC (Wong et.al, 2015). Analysis of forum messages is critical to instructors to obtain information about the quality of peer interaction taking place in the MOOC like a measure of learner engagement in higher-order thinking since the instructor has no face-to-face interaction with learners. Manual analysis of learners' posts is time consuming. Automatic classification can provide the requisite information to instructors enabling them to adopt measures to increase such type of interactions among learners in the remaining part or the next run of the course.

The existing work on message classifiers for the forum focuses on identifying messages for instructors that need their intervention. While they are useful, they do not provide a holistic view of the peer interaction happening in the forum. Hence, we developed a system that automatically classifies messages, helping the instructor capture and analyze the quality of peer interaction in DFs and the potential learning thereof. The forum data from two separate runs of an xMOOC, offered by the Indian Institute of Technology Bombay via IITBombayX platform, was used. The system automates the process of classification of forum messages into predefined classes that indicate the quality of peer interaction, based on the Interaction Analysis Model (IAM) (Gunawardena et.al, 1997). We explored multiple machine learning models such as decision tree- based models, rule-based models, statistical models, SVM, distance-based classifiers and deep learning models with the application of various pre-processing variants and analyzed their performances. The proposed system will be useful to instructors who have had to migrate to online teaching in the current Covid- 19 pandemic situation.

2. Related Work

We surveyed the existing literature on the importance of DF analysis and the machine learning based classifiers. Peer interaction fosters the ‘highest level of collaboration’ and critical reflection among learners (Toven- Lindsey et.al, 2015). MOOC forums often witness an overwhelming volume of orphaned posts that causes scatter (Manathunga et.al, 2017; McGuire, 2013), leading to cognitive overload for learners and limits peer interaction (Tawfik et.al, 2017).

Ntourmas et.al (2021) and Ntourmas et.al (2019) address the issue of information overload by developing supervised classification models to assist instructors in detecting forum discussions that need their intervention. Wise et.al (2016) addresses the issue of scatter and information overload in MOOC DFs by developing a model that categorizes and identifies threads based on their relation to the course content. In a similar work, Yi Cui and Alyssa Friend Wise (2015) investigated the extent to which the learner asked and the instructor answered content based questions in MOOC forums by building a classification model using linguistic features extracted by Lightside Researcher’s Workbench. Yiqiao Xu and Collin F. Lynch (2018) proposed an identification framework to identify and analyse help-seeking post.

The existing work on analysis of MOOC forum messages address the challenges instructors and learners face in locating the relevant set of messages from the overwhelming volume of forum messages and the resultant scatter. We found that the classification models that are built focuses on identifying messages posted by learners that instructors need to address. None of them aim to capture the quality of peer interaction that would aid instructors to estimate the amount of higher order thinking that is happening through the forum. This quality analysis is important feedback for instructors on the learning effectiveness of the course but it is extremely difficult to gauge manually. Hence, there is a need for automation of analysis of forum messages to measure the quality of peer interaction. We built a system that will automatically classify the messages in terms of the thinking level expressed in the messages. The system output will provide instructors with actionable insight into the extent of critical thinking occurring among the learners.

3. Methodology

In this section, we mention details about the dataset used and proposed system to be built in order to bridge the gap present in current work.

3.1 Context

We took the forum data from two separate runs of the same teacher professional development MOOC. We refer to the two runs as T1 (first run) and T2 (second run) in this paper. The teacher-learners were from multiple domains like Mathematics, English, Computer Science, Science and Social Science. The DF consists of Comment-thread (initialization of the post/discussion), Replies (responses to the comment thread) and Comments (responses to the replies) (Wong et.al, 2015). In this study, we consider each message as the unit of analysis, irrespective of their initiation source. Permission to use forum data for research purposes was obtained from the learners through IITBombayX.

3.2 Data Description

In this section we discuss data extraction, the thematic analysis and the nature of data observed.

3.2.1 Data Extraction

The prerequisite for the labelling phase was to extract out the relevant attributes from the data in the

T1 and T2 dataset based on a set of rules. The attributes are Comment_count, Comment thread_id, Parent_id and Sk_id. Comment_count is the number of comment replies in this thread. This includes all responses and replies, but does not include the original post that started the thread. Comment thread_id specifies the id the Comment Thread to which a specific Comment belongs. The Parent_id is the id of the response-level Comment that this Comment is a reply to. Sk_id if null, the type is Comment Thread, else type is Comment.

3.2.2 Thematic Analysis

The classification of our messages is based on the IAM (Gunawardena et.al, 1997). In our system, we followed the below classification:

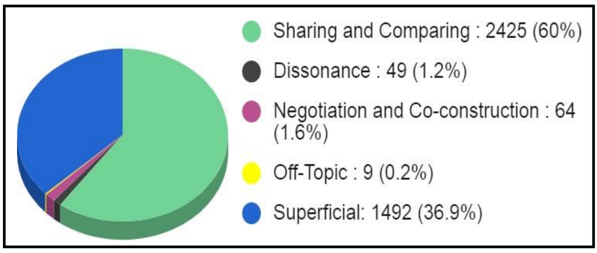
1. Superficial: Shallow messages with replies of greeting type from which learning is difficult to infer.
2. IAM phases: Messages that are elaborated in five phases (Gunawardena et.al, 1997).
    a. Sharing or Comparing of Information: Statement of observation or opinion, agreement or asking and answering questions to clarify details of statement, definition, description.
    b. Discovery of dissonance and inconsistency: Statements of disagreement, asking and answering questions to clarify the source and extent of disagreement.
    c. Negotiation of Meaning or Co-construction of knowledge: Clarification of the meaning of terms. Negotiation of the relative weight to be assigned to arguments, identification of areas of agreement or overlap among conflicting concepts.
    d. Testing and modification of proposed synthesis: Testing the proposed synthesis against ‘received fact’ as shared by the participants and/or their culture, testing against existing cognitive schema, testing against personal experience or formal data collected.
    e. Agreement/application of newly constructed meaning: Summarization of agreement(s), applications of new knowledge and metacognitive statements by the participants illustrating their (cognitive schema) has changed as a result of the interaction.
3. Off-topic: Content does not address any aspect of the current topic of discussion.

In this paper we referred to the above-mentioned a, b, c, as Sharing and Comparing, Dissonance, Negotiation and Co-construction. One of the authors did deductive coding of the T1 and T2 message rows using the mentioned categories as the codebook and following a set of general guidelines. For the Guided messages, labelling was done with respect to Focus Question. There were 6 Focus Questions in total, one for each week which were created by instructors to anchor discussions. The relevance of the Comment Threads is to be checked with respect to its Focus Question. For Non-Guided messages, since there are no focus questions, Comment Threads are to be judged considering the general nature of the sentence. Another author coded the same set of messages independently. Inter-rater reliability of the coding was established through discussion between the two coders till a complete consensus was achieved. There were 29355 data rows of T1 and 4707 data rows of T2. All 4707 of T2 data rowswere tagged. Since manual coding is labor intensive and time consuming, only 4039 data rows from T1 were arbitrarily picked and labelled.

3.2.3 Nature of Data

All messages including the comment threads, comments and replies were given equal importance and considered independently. The dataset contained two fields, namely message and category. Message refers to thread, comment, replies and the Category is obtained from coding the messages as mentioned in the above section. We proposed to build a model that aims to successfully classify the messages from the DF. We experimented with various classification approaches and compared them to check which would suit the best. We implemented machine learning algorithms and Keras Neural networks for categorizing the messages.

Superficial and Off-topic categories were not considered while building the classifier model since they are unlikely to contribute to learning. Only three phases of IAM were considered i.e., Sharing and Comparing, Dissonance, Negotiation and Co-construction because the remaining phases of the IAM were not found in the data. We found the data to be highly biased towards Sharing and Comparing with the majority of messages belonging to this phase. This is a common finding for MOOC forums where quality of learning has been found to be confined to the lowest level of critical thinking i.e., Sharing and Comparing information (Tawfik et.al, 2017). Consequently, we divided the IAM phases into two classes i.e., ‘Sharing and Comparing’ and ‘Higher order thinking (HOT)’ where HOT covers the IAM phases beyond ‘Sharing and Comparing’. For the forum being analyzed, HOT consisted of messages belonging to two such IAM phases – Dissonance, Negotiation and Co-construction. The resultant distribution among the categories is highlighted in the figure below.



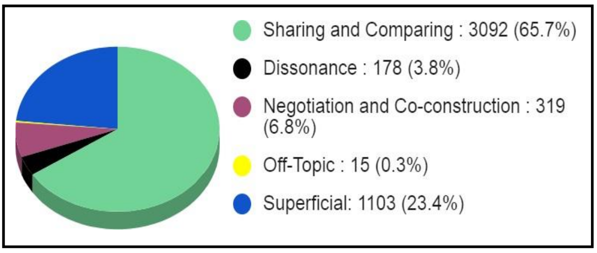


Figure 1. Distribution of Message Classes in T1. *Figure 2*. Distribution of Message Classes in T2.

3.3 Data Pre-processing

Our preprocessing consists of - tokenization, lowercasing, expansion of contractions, spell check, punctuation removal and custom stopword removal. In our work, we expanded contractions such as don’t and can’t to “do not” and “can not” in order to standardize the text. We created a list of custom stop words and removed them from the data. This is because, the predefined stop words in the English language nltk corpus contains words that are useful for our classification. Various combinations of the aforementioned preprocessing steps were carried out. The resultant data obtained in each case was fed into the models that were built. The three variants of pre-processing applied were - Variant 1: lowercasing, punctuation removal, Variant 2: lowercasing, punctuation removal, custom stopword removal and Variant 3: lowercasing, contraction removal, spell check, punctuation removal, custom stopword removal.

3.4 Proposed System

The below diagram illustrates the workflow that was followed for making the system.

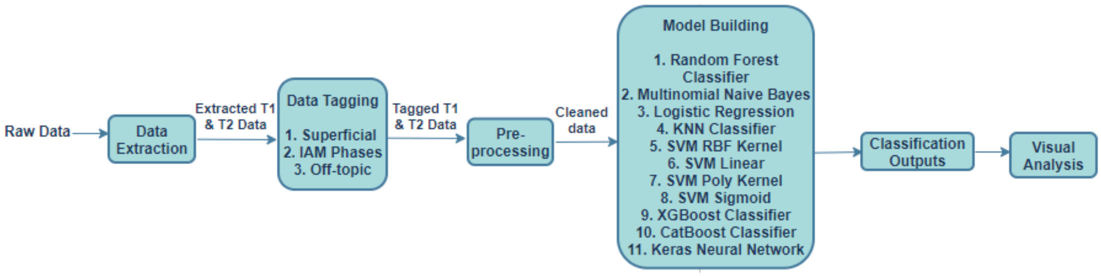


Figure 3. System Workflow.

4. Building and Training Models

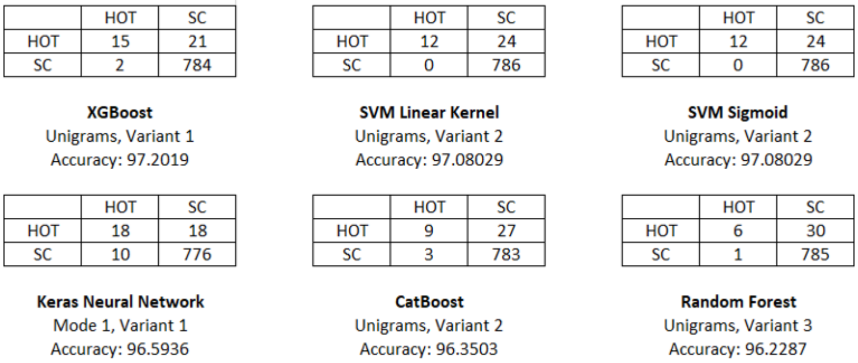
The labeled messages of T1 and T2 were combined and 80% of these messages were arbitrarily picked and used for training the models. The remaining messages were used for testing the models. The three pre-processing combinations i.e., Variant 1, Variant 2 and Variant 3 were applied on these messages which were then used to build models for automated classification of the messages.

The built models were Random Forest Classifier, Multinomial Naive Bayes, Logistic

Regression, KNN Classifier, SVM RBF Kernel, SVM Linear, SVM Poly Kernel, SVM Sigmoid, XGBoost Classifier, CatBoost Classifier, and Keras Neural Network. A pipeline approach was used for fitting the machine learning models on the training data. In the case of all these models, the variants of unigrams, bigrams and unigrams & bigrams both were implemented. Tf-idf was used as the feature to transform the input data where these variants were passed as parameters. In the case of Keras Neural network, the same two class classification approach was used i.e., labelling of messages as either 'Sharing & Comparing' or 'HOT'. The pre-processing steps mentioned in all the aforementioned variants were carried out on the data. Four modes were used to transform the words into features in the case of Keras Neural Network. They are binary, count, tf-idf and others.

5. Results and Discussion

The below shown figure displays the performance measure of the implemented classification models. All the types and variants for each model were implemented but a few were picked for illustration as shown below. For example, let's consider the output of the XGBoost Classifier using unigram and variant 1 of preprocessing. HOT refers to the Higher Order Thinking classes and SC refers to the Sharing and Comparing class. 15 HOT messages and 784 SC were correctly classified into their respective classes whereas 21 HOT messages were classified as SC and 2 SC messages were classified as HOT.



	HOT	SC
HOT	15	21
SC	2	784

XGBoost
Unigrams, Variant 1
Accuracy: 97.2019

	HOT	SC
HOT	12	24
SC	0	786

SVM Linear Kernel
Unigrams, Variant 2
Accuracy: 97.08029

	HOT	SC
HOT	12	24
SC	0	786

SVM Sigmoid
Unigrams, Variant 2
Accuracy: 97.08029

	HOT	SC
HOT	18	18
SC	10	776

Keras Neural Network
Mode 1, Variant 1
Accuracy: 96.5936

	HOT	SC
HOT	9	27
SC	3	783

CatBoost
Unigrams, Variant 2
Accuracy: 96.3503

	HOT	SC
HOT	6	30
SC	1	785

Random Forest
Unigrams, Variant 3
Accuracy: 96.2287

Figure 4. Performance Measure of Classification Models.

The general accuracy observed among the models was in the range of 95%-97%. Similar results were observed and no model outperformed the other by a great margin. As per our observations on the data in context, we found that the SVM Linear Kernel and XGBoost, both using unigrams, gave a relatively high accuracy of about 97% when it came to the overall classification of messages. Apart from SVM Linear Kernel and XGBoost, SVM Sigmoid Kernel model using unigrams, CatBoost using unigrams and Keras Neural Network gave satisfactory good results of around 96%. There were certain cases where the models were unable to classify any message belonging to the HOT category. This may be due to the disproportionate amount of data of the Sharing and Comparing class as compared to the HOT class. Models like KNN and Multinomial Naive Bayes were unable to classify 'Higher order thinking' type of messages correctly in all the considered combinations. We noticed a substantial increase in the accuracy on incrementing the training data that was fed to the models with the exception of some. SVM Linear Kernel and XGBoost gave the highest accuracy overall.

6. Conclusion

The objective of this study was to build an automatic classification system of MOOC forum messages to analyze the quality of peer interaction. We worked on the forum data from two separate runs of an xMOOC that ran on the IITBombayX platform for training and testing our system. We created a list of custom stop words and removed them from the data. Multiple machine learning models were explored and implemented in order to classify forum messages into predefined classes based on the IAM coding scheme. A general accuracy of 95%-97% was observed among the models.

Our system will enable instructors to gauge the learning effectiveness of their course by automatically analyzing and classifying the messages in the course forum into 'Sharing and comparing' and 'Higher order thinking'. The distribution of the messages into these two classes will provide instructors with an actionable insight into the extent of peer learning occurring in the forum. This insight is crucial as it gives an estimate of the higher order thinking ensuing in the form that is likely to lead to effective learning. In absence of face-to-face interaction in a MOOC course, such insight provides the instructors an estimate of the learning happening without having to wait for the assessment grades of the learners. It enables instructors to decide when to intervene to increase productive interactions among learners in the DF during the progress of the course.

Our classification model can be effectively generalized to smaller online courses. This is significant given the current covid-19 pandemic scenario where instructors have had to migrate to online teaching. As part of future work, we plan to build a rule-based architecture on top of the existing systemor apply deeper NLP techniques to process the messages. The system's efficiency could be increasedif the data was non biased or with more messages belonging to higher phases of IAM beyond 'Sharingand Comparing'.

Acknowledgements

Authors would like to acknowledge IITBombayX and Education Technology program at IIT Bombay for their support and permission to use their course data for this study and Nexteducation Research Lab (IITB) who were co-sponsor of these courses.

References

Cui, Y., & Wise, A. F. (2015, March). Identifying content-related threads in MOOC discussion forums. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (pp. 299-303).

Gunawardena, C. N., Lowe, C. A., & Anderson, T. (1997). Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of educational computing research*, *17*(4), 397-431.

Manathunga, K., Hernández-Leo, D., & Sharples, M. (2017, May). A Social learning space grid for MOOCs: exploring a FutureLearn case. In *European Conference on Massive Open Online Courses* (pp. 243-253). Springer, Cham.

McGuire, R. (2013). Building a sense of community in MOOCs. Campus Technology, August, 31-33. Retrieved April 2, 2014, from the Campus Technology

Ntourmas, A., Avouris, N., Daskalaki, S., & Dimitriadis, Y. (2019, July). Comparative study of two different MOOC forums posts classifiers: analysis and generalizability issues. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)* (pp. 1-8). IEEE.

Ntourmas, A., Daskalaki, S., Dimitriadis, Y., & Avouris, N. (2021). Classifying MOOC forum posts using corpora semantic similarities: a study on transferability across different courses. *Neural Computing and Applications*, 1-15.

Tawfik, A. A., Reeves, T. D., Stich, A. E., Gill, A., Hong, C., McDade, J., ... & Giabbanelli, P. J. (2017). The nature and level of learner–learner interaction in a chemistry massive open online course (MOOC). *Journal of Computing in Higher Education*, *29*(3), 411-431.

Toven-Lindsey, B., Rhoads, R. A., & Lozano, J. B. (2015). Virtually unlimited classrooms: Pedagogical practices in massive open online courses. *The internet and higher education*, *24*, 1-12.

Wise, A. F., Cui, Y., & Vytasek, J. (2016, April). Bringing order to chaos in MOOC discussion forums with content-related thread identification. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 188-197).

Wong, J. S., Pursel, B., Divinsky, A., & Jansen, B. J. (2015, March). An analysis of MOOC discussion forum interactions from the most active users. In International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (pp. 452-457). Springer, Cham.

Xu, Y., & Lynch, C. F. (2018). What do you want? Applying deep learning models to detect question topics in MOOC forum posts?. In *Wood-stock'18: ACM Symposium on Neural Gaze Detection* (pp. 1-6).

Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Evaluation of a Motion Capture and Virtual Reality Classroom for Secondary School Teacher Training

Sandra ALONSO[a*], **Daniel LÓPEZ**[a], **Andrés PUENTE**[a], **Alejandro ROMERO** [a], **Ibis M. ÁLVAREZ**[b] **& Borja MANERO**[a]
[a] *Universidad Complutense de Madrid, Spain*
[b]*Universitat Autónoma de Barcelona, Spain*
*sanalo05@ucm.es

Abstract: Nowadays having qualified and experienced teachers in school classrooms is considered to be of the highest priority in any society. However, most teachers report that they haven’t received sufficient practical training to manage disruptive situations in the classroom. Fortunately, virtual reality can provide a solution to this issue. This paper will introduce ClassroomVR-MotionCapture (CVR-MC) and its evaluation by experts. CVR-MC is an IT tool that can simulate a virtual classroom, thus allowing different users to face real-life problems that usually take place in real classrooms. The system captures the users’ tone of voice and the substance of their speech, as well as their gaze and corporal movements. Virtual students will react according to these parameters. To evaluate the usability and functionality of the tool, we conducted a study involving 14 education professionals. The research question that guided the study was the following: is it possible to use the CVR-MC system in teacher training to improve teachers’ communication skills for classroom climate management? The main conclusion of our study is that many participants described the CVRMC system as a friendly, safe and feasible environment for teacher training, especially for improving their classroom climate management competence. However, the study also found that the emotions detected through the users’ body expression did not match the emotions they reported feeling during the test.

Keywords: Non-verbal language, emotion recognition, virtual reality, teacher training, classroom climate, Secondary Education

1. Introduction

From an ecological perspective of teaching, classroom management creates, through the actions of teachers, an environment that facilitates achieving the learning goals and improving students’ socio-emotional well-being (Emmer & Stough, 2001). However, unexpected events may occur in the classroom that could disrupt instructional activities. The success of classroom management is dictated by the teacher’s competence in understanding and interpreting conflictive events in immediate circumstances.

Due to the importance of this issue, there is an urgent need to overcome teacher training deficiencies by improving teachers’ competence in classroom climate management. The Organization for Economic Co-operation and Development (OECD) and Teaching and Learning International Survey (TALIS) (Schleicher, 2020) state that Spain is well below the average for teachers who report having received training regarding management of students and classroom climate. Less than half of Spanish teachers (40%) report feeling prepared to control a class. This study also reveals that Spanish teachers spend the longest time trying to maintain order in class.

The lack of training in classroom management could potentially be resolved through the introduction of an initial training program to enhance preservice teachers’ communicative competence and increase their ability to respond to conflicts that frequently arise in secondary education classrooms.

Taking into account the deficiencies identified in current teacher training programs, particularly in relation to the competence for classroom climate management, our objective is to explore the usefulness of a virtual classroom system that has been created with the intention of developing this competence in initial teacher training.

1.1 Tools for Teacher Training

Our work is not the first approach to developing tools for improving teaching skills using virtual reality environments and technologies. Firstly, a US team developed the TeachLiveETM system (Barmaki & Hughes, 2015), a tool for analyzing non-verbal language. The authors conducted an experiment where only half of the participants received feedback on their non-verbal language and how to improve it during their lessons. They concluded that users who were given feedback experienced a significant improvement compared to those who did not receive it. Although this tool seems to have had good results, it did not consider the different elements of non-verbal language, such as gaze direction or voice tonality.

In Japan, Huang et al. (2016) developed a system for training non-experienced teachers in a virtual environment that simulated a high school class and used a Kinect device to register users' head and hand movements, as well as their voice. Overall, the participants had a good impression of the system, reporting that this virtual training system is needed. However, they pointed out that more animations were required, as well as an environment allowing the teacher to get closer to the students.

For its part, a study by the German Knowledge Media Research Center (Leibniz-Institut für Wissensmedien) (Sümer et al., 2021), validated a system that permits the evaluation of students' attention span through visible indicators of their level of participation in learning. Although the system proved to be efficient, it could be more effective using automated analysis.

All these systems provide very useful information about how non-verbal communication affects interactions with students. The information collected by these studies support the view that the systems are appropriate for improving competencies of non-experienced teachers. Moreover, the systems also allow teachers in training to have as many goes as needed to feel in command of the situation, which is the ultimate goal of the training exercise. However, we have not found any tool that can combine the detection and evaluation of non-verbal behavior (emotions and attitudes) with strategies to manage classroom conflicts.

Given the close relationship between both areas, our research attempts to provide a VR environment that would allow future teachers to develop their competence in managing classroom climate. We have called it CVR-MC, an extended work of which is called ClassRoomVR (Bocos Corredor et al., 2020). With this tool, we test users' competence by exposing them to three conflictive situations that are commonly faced in high school classrooms. this template, because they might have been overwritten by your local settings.

1.2 Non-Verbal Analysis Using Motion Capture

Due to the complexity of the field, the study of non-verbal language and how to capture it has been a long process. The behavioral aspect of emotion is reflected in facial expressions, vocal features, gestures and body postures. Thus, in a real situation, it is possible to analyze people's non-verbal language based on their natural movements (Torres, 2019). By observing body language, we can estimate the emotions people may be feeling (Ruano Arriagada, 2004).

If we add to the above the content of the message or even the user's biometric measurements, machine learning can help detect the emotions felt by the user in real time. For example, García-Magariño et al. (2019) put forward the Emopose tool, which is capable of analyzing an image projected on a 3D avatar and identifying basic emotions. This tool is based on the closest neighbor algorithm, choosing from its database the position closest to the one analyzed.

These studies use a Kinect device for detecting emotions through the user's body expression. They all agree that it is difficult to determine the user's predominant emotion just based on the person's posture. In sum, the use of virtual simulations is increasingly seen as an opportunity to provide pre-service teachers with unique opportunities to experience examples of classroom life in a controlled and structured manner. As stated McGarr (2021), it has been claimed that a complete psychological study is needed.

1.3 Research Question and Objective

This study seeks to answer the following research question (RQ): is it possible to use the CVR-MC system in teacher training to improve teachers' communication skills for classroom climate management? The main objective of this paper is to show our tool validation.

2. CVR-MC Architecture

The work by Bocos Corredor et al. (2020) established the basis of CVR-MC, which is aimed at practicing and improving teaching skills. Consequently, the authors created an environment that allows teachers to experience the climate and emotional factors present in the classroom. Based on this application, several extensions with new functionalities were developed, as shown below:

- Generation of conflictive situations and their corresponding management options: this allows situations to escalate and allows them to be easily generated.
- Tone of voice analysis: changes are detected during key moments of the simulation.
- Analysis of proxemia between interlocutors: distance between interlocutors (student and teacher) is analyzed and will determine the simulation response.
- Keyword detection in conflict management: keywords are classified to determine the response to the simulation and relate them to the user´s emotions.
- Possibility of choosing the execution platform: CVR-MC allows you to choose your preferred execution platform: a virtual reality environment or a desktop environment (PC).
- Analysis of the user´s body expression to establish the predominant emotion: the Perception Neuron and EmoPose system (García-Magariño et al., 2019) captured and analyzed the teacher´s body expression in order to associate it with an emotion.
- Storage of the most relevant parameters of the simulation: during simulation, key parameters are saved for analysis.
- Final feedback to the user after the simulation: final feedback is given to the user about his/her behavior and intention perceived during the simulation.

3. Evaluation of CVR-MC: Participants and Experimental Design

Fourteen participants, all experienced in the education field, were recruited to take part in the first CVRM test. Among them were high school principals, teachers and school counselors, lecturers of the Master's Degree in Training for Primary or Secondary Teachers and students of the Master's Degree in Teaching. The test took 25 minutes to complete and consisted of four stages:

1) *Test explanation*. Users were welcomed by a member of the technical team and were asked to sign the consent document to allow us to record audio and video for educational purposes.
2) *Test*. At this step, users were equipped with all the virtual reality immersion gear. When the user started to practice, a brief description of the environment was shown and the user was expected to simulate a response as if they were in a real classroom. After a short period of time, a disruptive situation would occur and the possible options that the user could choose to perform were shown. The player then had to simulate one of these actions in a natural way. The game recognized one of the paths taken and showed final feedback. Feedback consisted of three screens: 1) It indicated whether the action chosen was the most appropriate one, 2) It showed the main emotion captured by the system during the scene (fear, anger, joy, disgust, surprise or sadness), or 3) It reflected the variation in the tone of the player's voice during the interaction with the virtual students.

   Once the three scenes were completed, the devices were removed and the test recording stopped (see screenshot in Figure 1).

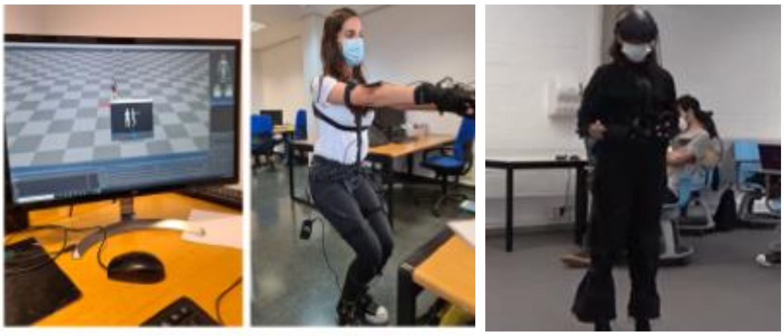


Figure 1. Perception Neuron Suit Calibration using Axis Neuron & Short Video Demonstration. https://youtu.be/r98GoEqEed0.

3) *Post-test Questionnaire*. We administered a questionnaire based on the Technology Acceptance Model (TAM, Davis, 1989). We adapted the questionnaire proposed by Huang et al. (2016), who conducted a similar study in Japan. Eighteen items grouped into three dimensions defined by the TAM were included.
   a) Perceived usefulness: how the user (teacher/teacher in training) perceives that the use of the learning environment (LE) can improve their competence to manage classroom conflicts (8 items). For example: "I felt that my behavior (movements, attitudes, words...) impacted students in the virtual environment."
   b) Perceived ease of use: if the user (teacher/teacher in training) perceives that the use of the LE will not involve any effort for them.
   c) Attitude towards app use: this refers to the emotions (positive or negative) experienced by the user (teacher/teacher in training) in their experience in the LE.
   d) Behavioral intention: how likely is that the user will use the LE as an environment to teach and learn about classroom climate management.

   To illustrate user answers, we used a seven-level Likert scale with the following correspondence: from total disagreement (1) to total agreement (7). For more information about the questionnaire, please contact the authors.

4) *Final interview*. The last stage consisted of a semi-structured interview (see next section) guided by a member of the research team. The interview lasted approximately 15 minutes and it explored the participants' evaluation of the experience, specifically focusing on the feedback provided about their emotions during the execution of the system test. In addition, participants were asked to write down suggestions for possible improvements to the tool.

4. Results

According to the model used to create the test (Davis, 1989), there are two main variables that affect users' acceptance and adoption of new technologies: perceived ease of use (1) and perceived usefulness (2). The answers to the questions in the post-test questionnaire (based on TAM), which users filed after the CVR-MC test, are shown in Table 1.

Table 1: *Summary of the Results Obtained in the Post-Test Questionnaire to Assess Users' Adaptation to CVR-MC (seven-level Likert-type scale)*

Criteria	Number of items	Cronbach's Alpha	Average	SD
Perceived usefulness (PU)	8	0.815	4.7	1.3
Perceived ease of use (UF)	4	0.840	5.1	1.3
Attitude towards the use of the learning environment (UA)	2	0.802	6.1	0.7

Behavioral intention (UI)	3	0.617	6.3	0.8

Questions around PU criteria showed information about the usefulness of CVR-MC, according to users. For this analysis, we added the responses obtained. In order to answer this question, a Cronbach's alpha was carried out, obtaining as a result an alpha of 0.915 and an average of 4.7 in responses. These responses show that users considered that the tool could improve their competence in managing conflicts in the classroom. This matches the positive reception shown by the users in the final interview and their perception of having had a safe environment for the practice of these skills.

Questions related to UF were analyzed with Cronbach's alpha, which produced an average of 5.1, showing that a large part of the users didn't have to make a big effort to use the tool; however, some reported feeling some discomfort. In the interviews, we were able to find out the limitations detected by the participants.

We then analyzed the questions looking at UA criteria (that is, those eliciting information about the attitude towards use), using a Cronbach's alpha that resulted in 0.802 and an average result of 6.1. These results showed that the user experience with the tool was relatively good. This data shows that although virtual reality environments are something new in this sector, they are also exciting and striking.

Finally, questions around UI criteria referred to the users' intention of using the tool. By means of Cronbach's alpha, we analyzed the results and obtained an alpha of 0.617 and an average of 6.3. With this result, and backed up by the interviews, we can conclude that, with certain improvements, the tool could play a part as a complementary use in teacher training:

"It seems a more significant formative experience than others I know because you can directly study a situation by doing a case study and thinking about what you would do. I find it much more interesting when you wear it, when you experience it, of course" (Participant 01).

5. Discussion

RQ: Is it possible to use the CVR-MC system in teacher training in order to improve teachers' communication skills for classroom climate management?

Yes, after analyzing the results we can affirm that our tool has potential to be used as a learning system that can contribute to teacher training. Users agreed that CVR-MC is a highly recommended support tool (4.7 on average) for teachers in training. CVR-MC will allow them to safely practice the necessary techniques to face a conflictive situation. In consequence, it will allow a greater development in their conflict resolution skills.

According to participants experts' comments, the feedback provided by CVR-MC is highly beneficial for trainee teachers (6.3 on average). Similarly, a joint discussion during teacher training can be very constructive and interesting for future teachers. However, we observed how some participants reported learning problems with the tool due to manageability, from which we can deduce that training prior to practical use would be helpful.

6. Conclusions and Future Work

This document presents the ClassroomVR-MotionCapture virtual reality tool that aims to improve the communication skills of secondary school teachers in order to improve conflict management within the classroom. In addition, the tool has been evaluated by experts in secondary education.

The main result we have obtained is that our tool, as confirmed by users, could be very useful in teacher training, as it would allow teachers to improve their skills in solving complex situations in the classroom. Participants highlighted the importance of the system's immersivity and its realism. Moreover, all participants recognized that the tool is a perfect complement to be introduced in the training of new teachers.

Our system sought to extract users' emotions through their corporal expression with a motion capture system. However, the main conclusion is that it is necessary to study how to track body expression in virtual environments. We concluded that, as the participants had to wear all the necessary devices to enter virtual reality (glasses, controllers and suit), their body expression changed. Consequently, their gestures were much more discreet than those done in real environments. In this sense, our system based on Emopose (García-Magariño et al., 2019), an application capable of inferring emotions through corporal expression, was not prepared for detecting the emotions felt by participants.

As future work, the first step will be to carry on with the study of body movement within virtual reality so that in the future we will be able to detect real emotions in virtual environments.

In addition, we will improve user experience by following some of the participants' suggestions. For instance, to improve the immersion experience we will replace the initial and final explanatory posters of the situation with a recorded audio.

Acknowledgements

This project has been funded by the Ministry of Science, Innovation and Universities of Spain (Didascalias, RTI2018-096401-A-I00).

References

Barmaki, R., & Hughes, C. E. (2015). Providing real-time feedback for student teachers in a virtual rehearsal environment. *Proceedings of ICMI '15: International Conference on Multimodal Interaction* (pp. 531–537). Seattle, WA, USA. https://doi.org/10.1145/2818346.2830604

Bocos-Corredor, M., Diaz-Nieto, A., Lopez-Garcia, A., Romero-Hernandez, A., Alvarez, I.M., & Manero, B. (2020). A VR game to improve communication skills in secondary-school teachers. *Proceedings of EDULEARN20. 12th International Conference on Education and New Learning Technologies* (pp. 8701–871). Online Conference. IATED Academy. . https://library.iated.org/view/BOCOSCORREDOR2020AVR

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Emmer, E. T., & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist, 36*(2), 103–112. https://doi.org/10.1207/S15326985EP3602_5

García-Magariño, I., Cerezo, E., Plaza, I., & Chittaro, L. (2019). A mobile application to report and detect 3D body emotional poses. *Expert Systems with Applications, 122*, 207–216. https://doi.org/10.1016/j.eswa.2019.01.021

Huang, H.-H., Ida, Y., Yamaguchi, K., & Kawagoe, K. (2016). *Development of a virtual classroom for high school teacher training.* Proceedings of Intelligent Virtual Agents. 16th International Conference, IVA 2016. (pp. 489–493). Volume 10011 of Lecture Notes in Computer Science, 2016. Springer, Cham. https://doi.org/10.1007/978-3-319-47665-0_61

McGarr, O. (2021). The use of virtual simulations in teacher education to develop pre-service teachers' behaviour and classroom management skills: implications for reflective practice. *Journal of Education for Teaching, 47*(2), 274-286. https://doi.org/10.1080/02607476.2020.1733398

Ruano Arriagada, M. T. (2004). *La influencia de la expresión corporal sobre las emociones: un estudio experimental*. (Doctoral dissertation). Retrieved from http://oa.upm.es/451/

Schleicher, A. (2020). *Teaching and Learning International Survey (TALIS) 2018. Insights and Interpretations*. OECD. Retrieved from https://www.oecd.org/education/talis/TALIS2018_insights_and_interpretations.pdf

Sümer, Ö., Goldberg, P., D'Mello, S., Gerjets, P., Trautwein, U., & Kasneci, E. (2021). Multimodal engagement analysis from facial videos in the classroom. *Computer Science. Computer Vision and Pattern Recognition*. *arXiv*. https://arxiv.org/abs/2101.04215

Sutton, R. E., Mudrey-Camino, R., & Knight, C. C. (2009). Teachers' emotion regulation and classroom management. Theory into Practice, 48(2), 130–137. https://doi.org/10.1080/00405840902776418

Torres, A. J. (2019). Interacción didáctica y lenguaje no verbal. Interconectando *Saberes, 7*. https://doi.org/10.25009/is.v0i7.2572

Identifying and Comparing Interaction Features of Different Topic Categories in Online Learning Discussions Supported by Danmaku

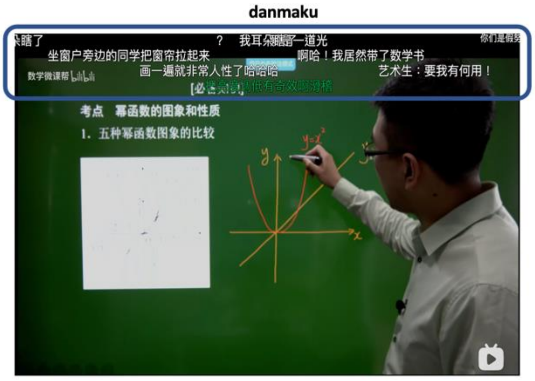
Bo YANG*
Department of Applied IT, University of Gothenburg, Sweden
*bo.yang@ait.gu.se

Abstract: With its interactive features, danmaku, a live-chat functionality allowing viewers to post messages right on the screen while watching videos, has numerous potentials of enhancing online learning interaction. In this paper, danmaku data generated by learners in one lecture of a high-school math course was retrieved. Coding based on content analysis was conducted to identify the interaction relationship, social network was then modeled, and results were compared between co-learner presence and idea exchange discussions. It was found that co-learner presence and idea exchange discussions showed differences in conversation structure, network topology and high-degree nodes. This study enhances the understanding of the interaction in MOOC learning facilitated by danmaku and provide evidence and basis for making use of danmaku discussions to better facilitate learning interaction.

Keywords: MOOCs, danmaku, online video learning, learning interaction, collaborative learning, content analysis, social network analysis

1. Introduction

Fig. 1 shows an example of danmaku, a system where user-generated messages are stored and displayed at their delivery timestamp along the MOOC video timeline and appear as moving subtitles on the screen when the video is played. This functionality originated in Japan in animation watching and got popular with teenagers and young people in Asia (Zhang & Cassany, 2019). Then, it was gradually introduced from animation watching to MOOC learning due to its interactive essence. It was found in an investigation that 76% of the students enjoyed this novel tool in their learning (Hu et al., 2017).



朵瞎了

？　我耳朵瞎了一道光

你们是假旁

坐窗户旁边的同学把窗帘拉起来

啊哈！我居然带了数学书

画一遍就非常人性了哈哈哈

艺术生：要我有何用！

Figure 1. A Screenshot of a Math MOOC Lecture Supporting Danmaku Functionality.

Over the recent years, social network analysis has been used in MOOC-learning research to reveal characteristics of interaction relationship (e.g., Joksimović, et al., 2016), analyze the text features and network characteristics (e.g., Dowell et al., 2015), or explore the correlation between interaction network and learning performance (e.g., Houston et al., 2017). However, these studies focused on traditional forum discussions instead of danmaku discussions. In addition, existing danmaku research mainly analyzed its usage in entertainment video and a few studies about learning began to explore the potential and effect of using danmaku in learning (e.g., Chen et al., 2019; Zhang et al., 2019). Although they confirmed the effectiveness of danmaku in supporting learning, a clear picture of the features and mechanisms of the learning interaction facilitated by danmaku have not been drawn yet.

To address this gap, this study aims to explore and identify the interaction feature and pattern implemented via danmaku in MOOC learning. Since previous studies found different topic categories (different sub-forums and learning content related or not) showed different patterns in MOOC-forum interaction and participation (Gillani & Eynon, 2014; Poquet & Dawson, 2016; Wise & Cui, 2018), it was hypothesized that similar effects also exist in MOOC-danmaku discussions but show different features as a "co-viewing" experience can be created by danmaku (Sun et al., 2018). Consequently, this study aims to answer the following research question:

RQ: What differences do the different topic categories show in conversation structure, network topology and core nodes in the learning discussion supported by danmaku?

2. Methods

2.1 Data Source

This study focused on a grade-10 math video lecture about quadratic function and data in this video was obtained from bilibili.com, a popular video portal. In total, 2,435 messages were collected and cleaned.

2.2 Tie Definition

In this study, interaction ties between learners through the danmaku messages in the video were defined into two categories: the directed tie where the message explicitly replied or referred to the preceding message (Joksimović et al., 2016; Kellogg et al., 2014) and the undirected tie where the learner simply sent relevant messages to participate in a group conversation of a certain topic (Jiang et al., 2014).

2.3 Content Analysis and Classification

Content analysis and subsequent manual coding were conducted and three coders familiar with danmaku participated in identifying and classifying the discussion topic and interaction tie.

First, we figured out the events triggered by video content as the context where danmaku discussions were situated since danmaku is one kind of event-based communication (Zhang & Cassany, 2020). Second, different topics of the danmaku discussions situated in those video-content events were located. Third, all the messages were checked in their relevance to each topic, the replying relationship and the temporal order based on their delivery timestamp. Thus, different rounds of discission achieved through danmaku were identified.

Two coders worked independently. After finishing coding, a third coder worked with the first two together to resolve the mismatch by using the majority-rule approach. Finally, the kappa coefficient of 0.76 and 0.88 was obtained for the classification of undirected and directed ties, respectively. Both values show substantial agreement (Viera & Garrett, 2005).

Although a variety of detailed topics were talked about in this lecture using danmaku, an obvious tendency was that many learners cared about the presence of others or themselves in the process of learning the lecture with the status of not seeing each other visually. The importance of social presence (Garrison, 1999) has been recognized by researchers in the formation and development of online learning groups (e.g., Kear, 2010). Of different elements composing social presence, co-presence was defined as the feelings of mutual awareness as well as inclusion and connection to a community (Biocca et al., 2003). Adopting this definition, we categorized the co-presence messages. In addition to the discussion about co-presence, learners also exchanged other general information. As a result, two general topic categories in all the messages were identified: 1) co-learning presence where learners tried to find and express the presence of others and themselves or their inclusion in a group of similar identities such as grade level, hometown and learning performance at school; 2) exchange of ideas in which knowledge information and emotional thoughts were communicated.

2.4 Network Modeling

The node list for the network was extracted from the danmaku message data. Although senders of the messages are all anonymous, each sender could still be represented by their encrypted user id. The edge list for the network was identified and mixture of directed and undirected edges was used based on the tie definition de-scribed in Section 2.2 and by manual classification in Section 2.3. Then, the two lists were imported into R and unweighted networks were modeled using the igraph package. Degree of all the nodes (the number of neighbors that a learner interacts with) were computed.

3. Results and Interpretation

3.1 Content Analysis

In total, of all the 2,435 danmaku messages, 1,203 and 315 messages achieved 44 and 57 rounds of undirected and directed discussions among learners, respectively while 917 messages were isolated expressions. Table 1 shows the structure of messages and rounds in the two discussion topic categories.

Table 1. *The Structure of Messages and Rounds in Each Discussion Topic Category*

	Co-learner presence	Exchange of ideas
Average word number in each message	4.7	10.8
Average message number in each round	7.57	20.3
n of messages in directed discussions	294	21
n of messages in undirected discussions	24	1,179
n of interaction rounds	42	59

According to the result, on average co-learner presence discussions usually used very limited number of short messages in each round while much larger number of longer messages appeared in exchange of ideas. This was probably resulted from another finding here that co-learner presence was mainly implemented by directed interactions and exchange of ideas was mostly achieved through undirected discussions, as Table 1 indicates. Since no function of replying to others' messages, such as the 'reply to' button, is provided in danmaku and all messages move across the screen and then disappear, learners need to use some techniques to explicitly respond to others' messages to complete the directed interaction and they have to do it quickly. In such scenarios, only part of all the viewers could make it and their words tended to be concise.

Besides, the topic of discussion for co-learner presence and idea exchange were different, which also explains the above-mentioned difference in message number and length. The starting message of each round in co-learner presence discussions often straightforward asked for co-learner presence information and could be easily provided by using simple words without extending the conversation. For example:

- Is there anybody watching?
- Yes.

Or

- I am a grade-12 student?
- I am also in grade-12.

On the contrary, starting messages in idea exchange discussions usually sought for help with understanding key knowledge points or emotion and attitude communication which could be more complicated, thus requiring extra number of longer messages. For example:

- How did the teacher get the 4x?

- Because the formula $(a+b)^2 = a^2+2ab+b^2$ should be used with “a” replaced by “x + 1” and “b” by “1”.

- I am a bit confused about this.
- $\{(x + 1) + 1\}^2 = (x + 1)^2 + 2(x + 1) + 1^2 = x^2 + 4x + 3$
- …

Or

- I am fond of this teacher’s style! He gave concise but understandable explanations.
- I cannot agree with you more. Of all the courses I attended, this free one is most effective.
- Same with what I felt!
- Is he really so good? Maybe I can only know it if I pass the exam.
- …

3.2 *Social Network Analysis*

As Fig. 2 indicates, the co-learner presence network (see Figure. 2a) consists of many independent small sub-networks while the idea exchange network (see Figure. 2b) is basically a large inter-connected network. This difference between the independence and interconnection depends on the number of shared nodes bridging different sub-networks (rounds of discussions) and may indicate that learners who joined the co-learner presence discussion participated in only few rounds of discussions whereas one learner in idea exchange discussions could took part in much more different rounds.

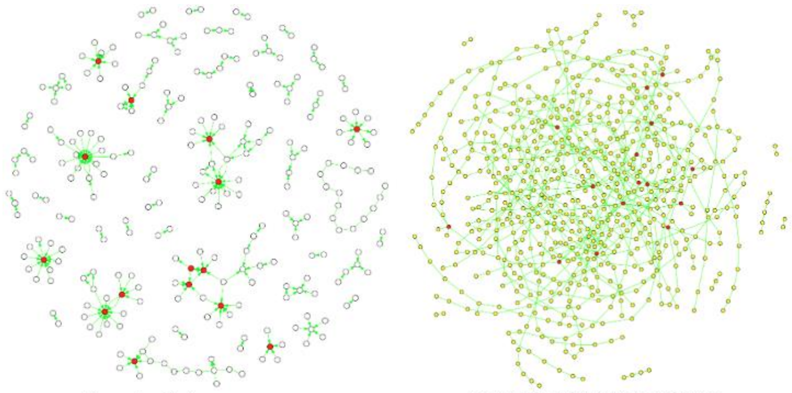


Figure.2a. Co-learner presence

Figure.2b. Information exchange

Figure 2. Social Networks Constructed for the Co-learner Presence and Idea Exchange Interactions.

The top 15 learners with high node degree in each network were highlighted with red in two networks, respectively. Table 2 reports the number of starting messages and percentage of other nodes connected through the starting message for the top 15 learners ranked by node degree in the two networks. In the co-learner presence network, it could be found that these high-degree nodes are cores of the small sub-network as most of them were starting message sender and therefore interacted with other learners who replied. This goes in consistency with the finding in Section 3.1 as very limited number of messages appeared in each round of discussion in co-learner presence interaction, thus achieving those independent small sub-networks. In total, 29 distinct learners appeared in the top 15 lists for the co-learner presence and idea exchange networks. Except for user “167dcd56” (highlighted in orange in Table 2) who stayed as the high-degree learner in both networks, the other 28 learners had high degree in only one of the two networks. It can be figured out that generally top learners in the two networks were different and that getting well connected in one network did not necessarily ensure good connectivity in the other. This is similar to the finding of Wise and Cui (2018) that high-degree MOOC learners in the content-related forum discussions were largely different from those in non-content discussions.

Table 2. *Top 15 Learners with the Number of Starting Messages They Sent and Percentage of Other Nodes Connected through the Starting Message in All the Connected Nodes of the Learner*

Rank	Co-learner presence			Idea exchange		
	User ID	①	②	User ID	①	②
1	cd7e6fec	2	73%	cbd6d772	2	33%
2	d4160a6a	1	100%	87a33001	0	0%
3	*167dcd56*	2	87%	ff6b175a	0	0%
4	a7fa6872	1	60%	19bbfa53	0	0%
5	875320b	1	100%	651617da	0	0%
6	aaa4e453	2	50%	8aabf982	0	0%
7	c24718f6	1	50%	49079dc1	0	0%
8	8a45564e	1	100%	*167dcd56*	0	0%
9	a7ae6d97	1	57%	3045389f	0	0%
10	72d5b521	1	43%	3e1c35ea	1	40%
11	220d53f8	0	0%	5ab1b637	0	0%
12	2c5749fd	1	100%	1156fbf6	0	0%
13	3e8f3c7f	1	100%	155628fd	0	0%
14	7bec74e8	1	83%	1bb84137	0	0%
15	4b9051ab	1	100%	2f4abb28	0	0%

① refers to the number of starting messages sent by the learner in any round of discussions
② refers to the percentage of nodes connected by the starting messages in all the connected nodes of the learner.

In addition, for 13 out of the 15 high-degree nodes in the co-learner presence network, at least half of all the other nodes connected with them were tied through the starting message. In contrast, only two out of the 15 learners sent starting messages and most of the ties between them and their connected nodes were not implemented through the starting message. This shows that sending starting messages (e.g., asking if co-learners exist or sharing the date of watching the lecture) which can initiate a round of discussion could help achieve interaction with other learners in co-learner presence network. This can be probably explained by the herding effect in danmaku since viewers could be easily affected by observing others' danmaku messages and leading messages stimulated subsequent ones (He et al., 2017). However, how connectivity formed in the idea exchange discussions still requires further research.

4. Conclusion

This is a very timely study as the whole world is still in pandemic now and online learning are playing the most crucial role it ever has. Learning videos supporting danmaku are getting popularity in MOOCs and danmaku is serving as a tool for achieving learning interaction. This study offered insight into how learners interacted differently in the co-learner presence and idea exchange discussions in a math MOOC lecture and the resultant findings in topic categories, conversation structure, network topology and core nodes could shed light on how to make full use of danmaku to improve MOOC learning interaction between learners. For example, as many learners used danmaku for co-learner presence interaction, the course instructor can intentionally harness social presence cues during the lecture to help learners engage more with their peer learners. In addition, since interaction on co-learner presence and idea exchange showed rather different patterns, MOOC portals can adjust the interface and functionality accordingly to better meet different needs of learners in online video learning.

Besides, Leng et al. (2016) found that danmaku video did improve students' learning outcomes in a small-sample eye-gaze experiment whereas how danmaku helps learners implement interaction needs more research, especially in the non-experimental environment. This paper identified the active effect of danmaku on conducting the co-presence and idea exchange interaction in an online math lecture. Since previous studies found interactions correlate with learning experience and performance (e.g., Houston et al., 2017), findings here could suggest the potential of danmaku in influencing both the experience and outcome.

However, this is a pilot study about one math lecture and is limited to a small facet of the danmaku-supported learning interaction. Further research is necessary to present more aspects of how danmaku shapes learners' communication and influences their learning performance. For example, lots of isolated messages which failed to realize interaction with others were identified in this research and they will be studied later to explore the reason for the failure.

References

Biocca, F., Harms, C., & Burgoon, J. K. (2003). Toward a more robust theory and measure of social presence: *Review and suggested criteria. Presence, 12*(5), 456–480.

Dowell, N., Skrypnyk, O., Joksimović, S., Graesser, A. C., Dawson, S., Gašević, D., et al. (2015). Modeling learners' social centrality and performance through language and discourse. Proceedings of the 8th international conference on educational data mining (pp. 250–257). New York, NY, USA: ACM.

Garrison, D., Anderson, T., & Archer, W. (1999). Critical Inquiry in a Text-Based Environment: Computer Conferencing in Higher Education. *The Internet and Higher Education, 2*(2-3), 87–105.

Gillani, N., & Eynon, R. (2014). Communication patterns in massively open online courses. *The Internet and Higher Education, 23*, 18–26.

Houston, S. L., Brady, K., Narasimham, G., & Fisher, D. (2017). Pass the idea please: The relationship between network position, direct engagement, and course performance in MOOCs. Proceedings of the 4th (2017) ACM conference on learning@ scale (pp. 295–298). New York, NY, USA: ACM.

He, M., Ge, Y., Chen, E. H., Liu, Q & Wang, X. S. (2017). Exploring the emerging type of comment for online videos: danmu. ACM Transactions on the Web (12)1.

Hu, Y., Hao, Q., Zhou, Y., & Huang, Y. (2017). Interactive teaching and learning with smart phone app in Optoelectronic Instrument course. 14th Conference on Education and Training in Optics and Photonics, ETOP 2017. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series.

Jiang, S., Fitzhugh, S. M., & Warschauer, M. (2014). Social positioning and performance in MOOCs. Proceedings of graph-based educational data mining workshop at the 7th international conference on educational data mining (pp. 55–58). CEUR-WS.

Joksimović, S., Manataki, A., Gašević, D., Dawson, S., Kovanović, V., Kereki, D., et al. (2016). Translating network position into performance: Importance of centrality in different network configurations. Proceedings of the 6th international conference on learning analytics & knowledge (pp. 314–323). NY, USA: ACM New York.

Kear, K. (2010). Social presence in online learning communities. In Proceedings of the 7th International Conference on Networked Learning (NLC2010), pp. 541–548.

Kellogg, S., Booth, S., & Oliver, K. (2014). A social network perspective on peer supported learning in MOOCs for educators. *International Review of Research in Open and Distance Learning, 15*(5). http://doi.org/10.19173/irrodl.v15i5.1852.

Leng, J., Zhu, J., Wang, X., & Gu, X. (2016). Identifying the Potential of Danmaku Video from Eye Gaze Data. 2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT), pp. 288-292. https://doi: 10.1109/ICALT.2016.155.

Poquet, L., & Dawson, S. (2016). Untangling MOOC learner networks. Proceedings of the 6th international conference on learning analytics & knowledge (pp. 208–212). NY, USA: ACM New York. http://doi.org/10.1145/2883851.2883919.

Sun, Z., Sun, M., Cao, N., & Ma, X. VideoForest: interactive visual summarization of video streams based on danmu data. SIGGRAPH ASIA 2016 Symposium on Visualization. ACM (2016).

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine, 37*(5), 360-363.

Wise, A. F., & Cui, Y. (2018). Learning communities in the crowd: characteristics of content related interactions and social relationships in mooc discussion forums. *Computers & Education, 122*, 221-242.

Yue Chen, Qin Gao, Quan Yuan & Yuanli Tang (2019) Facilitating Students' Interaction in MOOCs through Timeline-Anchored Discussion. *International Journal of Human–Computer Interaction, 35*(19), 1781-1799, DOI: 10.1080/10447318.2019.1574056

Zhang, Y. B., Qian, A. P., Pi, Z. L., & Yang, J. M. (2019). Danmaku related to video content facilitates learning. *Journal of Educational Technology Sys-tems, 47*(3), 359–372.

Zhang, L., & Cassany, D. (2019). The 'danmu' phenomenon and media participation: Intercultural understanding and language learning through 'The Ministry of Time'. *COMUNICAR, 27*(58), 19-29.

Zhang, L., & Cassany, D. (2020) "Making sense of danmu: Coherence in massive anonymous chats on Bilibili. com." *Discourse Studies, 22*(4), 483-502.

Explore the Contribution of Learning Style for Predicting Learning Achievement and Its Relationship with Reading Learning Behaviors

Fuzheng ZHAO[a], Bo JIANG[b], Juan ZHOU[c] & Chengjiu YIN[d*]
[a]*Graduate School of System Informatics, Kobe University, Japan*
[b]*Department of Educational Information Technology, East China Normal University, China*
[c]*School of Environment and Society, Tokyo Institute of Technology, Japan*
[d]*Information Science and Technology Center, Kobe University, Japan*
*yin@lion.kobe-u.ac.jp

Abstract: Prediction is an important branch of research in learning analytics, in which the prediction of learning achievement has much practical value for improving instructional management and enhancing learning effectiveness. As a type of cognitive data, students' learning style data offers great potential for predicting their learning achievement. Based on the analysis of the contribution of learning style data on prediction model creation, this study uses the Felder and Silverman learning style scale to examine 238 students' learning styles as feature elements and explores the feature importance using six machine learning algorithms to create models for learning achievement prediction. Besides, to identify the relationship between learning styles and learning behaviors, and the hidden learning patterns behind learning styles, the study collected reading log data using the E-book system for correlation and principal component analysis. It was found that the Decision Tree model obtained the best results in terms of accuracy and other indicators. Secondly, the VisualScore feature showed the greatest influence on all the six models used. Thirdly, the study also found that learning styles were highly correlated with repeated learning and marking behavior in reading behavior. Finally, the analysis showed that the visual and verbal dimensions under the VisualScore features had three common learning patterns of repeated reading, marking, and mobile reading, in addition to differences in learning patterns in terms of time spent.

Keywords: Learning style, learning prediction, reading learning behavior

1. Introduction

In the e-publication era, learning analytics provides a huge analysis and mining potential (Zhao, Huang, & Yin, 2018) to rethink the role (Yin et al., 2019) and strategy of education technology in the learning practice (Yin & Huang, 2018). Various predicted methods were selected to detect whether they have the ease and effectiveness of predicting effects (Huang et al., 2020). Learning style is mainly expressed by preference in learning methods. It is worth noting that learning style has a certain correlation with cognitive ability, but it has no absolute relationship with the strength or weakness of ability (Hames & Baker, 2015).

This study experimented to collect the data by the Index of Learning Styles and reading learning behavior by E-book system, to understand the contribution of learning style data to predicting achievement, figure out the most optimal features, as well as explore the relationship between learning style data and reading learning behavior.

2. Methodology

The learning style data including four variables, such as ActiveScore, SensingScore, VisualScore, and SequentialScore, was collected from 238 participants by the Index of Learning Styles questionnaire. In this study, the Scikit-learn was used to make model creation, such as Decision Tree (DT), Random

Forests (RF), XGBoost (XG), Logistic Regression (LR), Support Vector Machines (SVM), and K-nearest Neighbors (KNN). Subsequently, as a binary classification case, Accuracy, F-score, Recall, Precision, and AUC meet the requirements of evaluation for model performance. Finally, we used impurity-based feature importance, coefficients feature importance, and permutation feature importance to calculate feature importance.

To analyze the relationship between reading learning behavior and VisualScore. 238 college students were recruited and used an E-book system to reading a learning material. The learning reading behavior consists of 11 basic reading variables, such as PC, Mobile, Tablet, Bookmake, Memo, Highlight, Underline, Prev, Next, Readtime, and Readpage. We used three groups, including Information Gain (IG) and Gain Ratio (GR), ANOVA and χ^2, and Fast Correlation Based Filter (FCBF), to perform correlation analysis. In addition, the principal component analysis (PCA) was used for exploring the learning patterns behind learning styles.

3. Results

3.1 Model Performance and Feature Importance

The prediction performance of the six prediction models is based on the learning style data in Accuracy, Precision, Recall, F1-score, and Auc five metrics. On the whole, the DT model has an average score above 0.7. In contrast, the prediction performance of the other five models is unevenly distributed across the five indicators, and the scores fluctuate relatively widely.

In terms of feature importance, it is found that there is a nearly similar proportion in the contribution of features to prediction performance. Although 5 models except DT do not meet the basic requirement of good prediction performance, the common thing is that the VisualScore feature offers the most important contribution for prediction with the largest proportion (0.752 for DT, 0.240 for RF, 0.240 for XG, -0.459 for SVM, -0.046 for LR, 0.159 for KNN).

3.2 Correlation between Learning Style and Reading Learning Behavior

The study analyzed the relationship in terms of the amount of information, sample variability, and inter-sample distance respectively. It was found that the first correlation exists between the Prev (IF 0.023, GR 0.011) behavior, which refer to scrolling back to read the material, and the VisualScore learning style. Second, Memo (ANOVA 8.501, χ^2 4.497) and HighLight (ANOVA 5.909, χ^2 3.199), also called mark behavior, are highly correlated with learning style. Specifically, from the IG, GR, and FCBF indicators, the Prev feature occupies first place, with scores of 0.023, 0.0115, and 0.017 respectively.

3.3 Learning Patterns behind Learning Style

The study used principal component analysis to extract common factors on students' reading behavior and then organized them into mutually independent categories. First, for the analysis of learning patterns of students with visual learning tendency, their reading learning behaviors were analyzed by PCA, and the 11 features were extracted as principal components to form four common factors, which are time-spending category (Next 0.94, Readtime 0.927, and Readpage 0.9), nark category (Highlight 0.889, Memo0.784 and UnderLine 0.631), repeated reading category (Prev 0.659 and PC 0.665), and mobile device category (Tablet 0.779 and UnderLine 0.297)

Second, the learning pattern of students with verbal learning tendencies was determined. There are three categories of learning patterns. The first one is a repeated reading category (Prev 0.951, Readpage 0.913, and Readtime 0.819), followed by the second category, mark category (Memo 0.933 and Highlight 0.868), finally, mobile device category (Mobile 0.633 and Underline 0.441).

4. Conclusion

Prediction model selection and feature importance. In terms of binary classifications that based on learning style data, DT model performs best. For the prediction performance, it was found that the DT model outperforms other models, with an average score of above 0.7. This result not only provides an insight of impacts and contribution of learning style data to predict students' achievement, but also figures out the extent to which various categories of learning styles affect the students' achievement, as well as ranks the importance of learning styles.

Regarding the feature importance, it is obvious that the VisualScore feature contributes most to the model construction, no matter what kind of models are based on various algorithms. However, the largest proportion of contribution exhibited by the VisualScore feature occurs in the DT model. There are two points worth noting when performing feature importance calculations. First, the feature importance calculation uses different calculation methods, and this diversity phenomenon is determined by the algorithm behind the model. Second, in terms of feature importance some models are limited by the characteristics of the algorithm behind them, and tend to equalize the feature contributions during model crating, such as the KNN model.

The results obtained from multiple correlation calculation methods are better than the analysis results under a single dimension. This study used several different correlation analyses between learning styles and reading behaviors. The analysis revealed that the two behaviors, repeated reading and marking, are highly correlated with learning styles. The former was explored from the perspective of informativeness, concentrating on the informativeness and uniqueness of the behavior to the learning style representation. The latter puts the analytical perspective on the variability of the sample group and within-group, emphasizing the variability between learning styles and reading behaviors.

Learning patterns behind learning style. For students who tend to be visual or verbal, repeated reading and marking learning methods are the most important learning mode. Based on the results of the PCA analysis, it was found that, on the one hand, verbal and visual learning styles have similar learning patterns, such as repeated reading, marking, and using mobile devices. On the other hand, they show differences the time-spending. In particular, students with visual-prone learning styles show the time-spending pattern. However, this difference has not been further confirmed and there is a lack of understanding of why it occurs, which will be a focus of research in the future.

Acknowledgements

Part of this research work was supported by the Grants-in-Aid for Scientific Research Nos. 21H00905 and 18H04125 from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan. We would like to thank all the people who prepared and revised previous versions of this document.

References

Hames, E., & Baker, M. (2015). A study of the relationship between learning styles and cognitive abilities in engineering students. *European Journal of Engineering Education*, *40*(2), 167-185.

Huang, A. Y., Lu, O. H., Huang, J. C., Yin, C. J., & Yang, S. J. (2020). Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs. *Interactive Learning Environments*, *28*(2), 206-230.

Yin, C., & Hwang, G. J. (2018). Roles and strategies of learning analytics in the e-publication era. *Knowledge Management & E-Learning: An International Journal*, *10*(4), 455-468.

Yin, C., Yamada, M., Oi, M., Shimada, A., Okubo, F., Kojima, K., & Ogata, H. (2019). Exploring the relationships between reading behavior patterns and learning outcomes based on log data from e-books: A human Factor Approach. *International Journal of Human–Computer Interaction*, *35*(4-5), 313-322.

Zhao, F., Hwang, G. J., & Yin, C. (2021). A Result Confirmation-based Learning Behavior Analysis Framework for Exploring the Hidden Reasons behind Patterns and Strategies. *Educational Technology & Society*, *24*(1), 138-151.

The Acceptance of Mobile Games to Improve Filipino and English Vocabulary among Children from Urban and Rural Areas

May Marie P. TALANDRON-FELIPE[ab*] **& Ma. Mercedes T. RODRIGO**[a]
[a]*Ateneo Laboratory for the Learning Sciences, Ateneo de Manila University, Philippines*
[b]*University of Science and Technology of Southern Philippines, Philippines*
*maymarie.talandron-felipe@ustp.edu.ph

Abstract: The Programme for International Student Assessment (PISA) 2018 National Report of the Philippines showed that students from the southern part of the country with a mother tongue other than English or Filipino had lower proficiency levels in reading than students from the National Capital Region. It was also reported that students residing in urban communities outperformed those from rural areas. This paper first investigates the attitude and perception towards Filipino and English among elementary students from urban and rural areas in the southern region of the Philippines who are non-native speakers of Filipino and English. Then they were introduced to mobile-assisted language learning through the *Ibigkas!* educational mobile games which were developed to help improve the students' Filipino and English vocabulary. Results showed that there are differences in the perception and usage of Filipino and English between the groups and these are consistent for both languages. With regards to the *Ibigkas!* games, the positive feedback of the students in terms of the game-based learning engagement and intrinsic motivation inventory as well as some learning gains from their first time playing them are regarded as indications of acceptance and the promising potential of the said mobile games.

Keywords: Mobile-assisted language learning, game-based learning, Filipino language learners, mother tongue-based multilingual education, second language learning

1. Introduction

The Philippines is one of the most linguistically diverse nations in South East Asia with around 187 languages spoken across different regions, resulting into multiple mother tongues (MT) (Metila, Pradilla, & Williams, 2016). Although many of these languages exhibit dialectal variation, it is important to note that these are considered languages, not dialects. This means that these variations are so different that a speaker of one language may not easily understand communication in the other languages (McFarland, 2008). This diversity is associated with the country's geographical composition: an archipelago of 7,100 islands divided into three major groups. The northern group of islands which include the National Capital Region (NCR) is named Luzon, the central group is Visayas, and the southern group is Mindanao. The majority of people from NCR speak Tagalog while the rest of Luzon mostly use Ilokano, Bicol, Kapampangan, and Pangasinan. In the Visayas, the dominant languages are Cebuano, Hiligaynon, and Waray. In Mindanao, the use of the 8 mentioned languages is more dispersed (Jubilado, 2004).

In the education sector, the Vernacular Education Policy (1957) was implemented to use various regional lingua francas (LF) and English as the language-in-education but was replaced by the Bilingual Education Policy (BEP) in 1970 that required the use of Filipino (a standardized version of Tagalog) and English. Moreover, both the 1973 and 1987 Philippine Constitutions formally designated Filipino as the national language and symbol of national identity and culture. This further strengthened the implementation of BEP that ensures the development of literacy in Filipino as a linguistic symbol which is a representation of our unity and the literacy in English as the perceived universal language in preparing globally-competitive students. Under the BEP, Filipino and English were the primary mediums of instructions in schools across grade levels and linguistic regions (Dekker & Young, 2005).

In 2009, this policy changed. The Philippine Department of Education (DepEd) established Mother Tongue-Based Multilingual Education (MTB-MLE) (Philippines Department of Education, 2009) that aimed to have children start formal education in the language they know best. This decision was aligned with prior research that showed that the students' mother tongue offered the best foundation for learning additional languages and academic content (Barron, 2012; World Bank, 2005). Learning was therefore best achieved through mother tongue immersion in the first few years of formal education (Ball, 2010). Since its nationwide implementation in all public schools in 2012, the local mother tongue is taught as a language subject and used as medium of instruction in teaching all subjects from kindergarten to third grade (K3). Starting from fourth grade (K4), the medium of instruction shifts to Filipino for *Araling Panlipunan* (Social Studies), *Edukasyong Pantahanan at Pangkabuhayan* (Livelihood Education), *Musika, Sining, Edukasyong Pangkatawan at Pangkalusugan* (Music, Arts, Physical Education, and Health), and *Edukasyon sa Pagpapakatao* (Values Education). For Science and Mathematics, the medium of instruction shifts to English (Barnachea, 2013).

Although both Filipino and English are taught as language subjects from K1 to K3, the shift in the medium of instruction imposes challenges to students. At this point, they may not have yet mastered even their mother tongues, much more so Filipino and English which are essential in acquiring new knowledge and skills from K4 onwards. Also, as Tagalog is the base of Filipino, it may follow that Filipino is already considered the medium of instruction for Tagalog-speaking students from K1 to K3 and they only need to adjust to the shift to English for some subjects when they reach the fourth grade. This results to a disadvantage for students whose mother tongue is not Tagalog for they still have to learn Filipino as their second language and English as their third language in order to adjust to the shift in the medium of instruction (Tupas & Lorente, 2014).

The consequences of these policies on learners are partially reflected in the results of international tests. The Programme for International Student Assessment (PISA) 2018 National Report of the Philippines showed that the country obtained an average of 340 points in Overall Reading Literacy and was classified at Proficiency Level 1a, two levels lower in comparison to the average of the Organization for Economic Cooperation and Development (OECD) countries which was 487 and ranked at Proficiency Level 3. Within the Philippines, 94.70% from Region 12 (Central Mindanao) obtained lower than proficiency level 2, including the 33.97% with proficiency level 1c and below. On the other hand, 32.98% from the National Capital Region obtained proficiency levels 2 to 4 while they only have 8.56% with proficiency level 1c and below. It is also interesting to note that in the PISA report, the type of community was also seen as a factor on the significant differences on the average performance of students. Students residing in urban communities outperformed those from rural areas in reading literacy (355 against 313 points) (Philippine Department of Education, 2019).

Do students value English and Filipino language learning? Do they see these skills as beneficial or important? Among non-native speakers of English and Filipino, we hypothesize that: (1) urban-based students accept and use the languages more and (2) urban-based students have better language skills than rural students.

The study reported in this paper was conducted with the following goals: (1) to examine the differences in the usage, attitude, and perception towards Filipino and English among elementary students from urban and rural areas in southern Philippines, (2) to investigate the potential of educational mobile games to help improve the students' Filipino and English vocabulary, and (3) to compare the performance of urban-based and rural-based students in a digital game for English and Filipino vocabulary.

2. Mobile Games for Language Learning

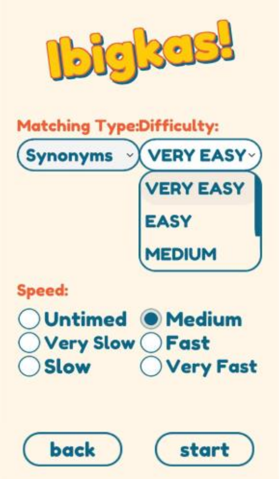
2.1 Digital Games and Mobile-Assisted Language Learning (MALL)

Advances in mobile devices, particularly on handheld gadgets such as personal digital assistants (PDAs), tablets, and smart phones have enabled the use of multimedia in mobile applications and provided both educators and learners a wide variety of learning resources (Huang, Chen, & Chen, 2009). The development of such diverse educational mobile applications added value to how mobile technologies can support learning outside the classroom (Mouza & Barrett-Greenly, 2015). According to Ozdamli and Cavus (2011), mobile learning has seven core characteristics: (1) ubiquitous because it provides more spontaneity than other learning models, (2) portable because mobile learning tools are relatively small and do not require much space and complicated set ups, (3) blended because it can be combined with other modes of instructions, (4) private because most mobile devices have a one-to-one ratio which means only one learner at a time has access to the resource and can independently work on the tasks, (5) interactive because m-learning environments make use of latest multimedia technologies in developing engaging tasks, (6) collaborative because mobile technologies support communication between students and with the teachers, and (7) instant because mobile technologies allow the delivery of immediate queries and responses. Developers leverage these characteristics in creating applications that can be used for both formal and non-formal education (Kukulska-Hulme, Lee, & Norris, 2017).

Over the last decade, there has been a significant growth in research work on digital games and language learning (e.g. Cornillie, Thorne, & Desmet, 2012; Godwin-Jones, 2014; Reinders, 2012; Sykes, 2018). Coupled with the increasing popularity of mobile technologies for learning, digital games have also undergone a rapid shift to mobile platforms (Giannakas et al., 2018) and as such have been utilized for mobile-assisted language learning (MALL), an area within mobile learning that focuses on various language learning (Miangah & Nezarat, 2012). MALL research has shown that mobile devices can indeed be effective tools for delivering language learning materials to the students and it allows them to autonomously study a second language (Kukulska-Hulme et al., 2017). Educators have used games in the MALL context which have been shown to improve language skills such as listening, vocabulary, and grammatical accuracy to help improve language proficiency and integrate them into the curriculum as supplementary activities (Sykes, 2017). One example of such a game is *Ibigkas!* (Rodrigo et al., 2019).

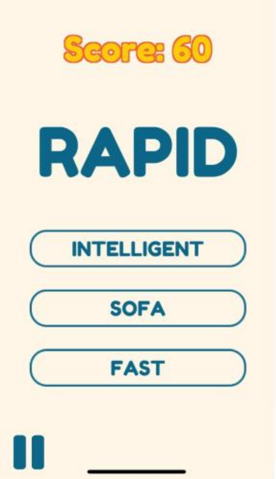
2.2 The Ibigkas! Games

Ibigkas! (translated as "Speak Up!" in English) is a mobile game developed to help improve English literacy skills (Rodrigo et al., 2019). It focuses on the recognition of English words, specifically students' knowledge of rhymes, synonyms, and antonyms. It is a drill-type game that can be played in single player or multiplayer modes. In a single player mode, the player starts by selecting the type of game, the level of difficulty, and speed (Figure 1a). The game gives the player a target word and the player must choose from the 3 options the correct rhyme, synonym, or antonyms depending the game s/he is playing (Figure 1b).



Matching Type: Synonyms
Difficulty: VERY EASY
Speed:
○ Untimed ● Medium
○ Very Slow ○ Fast
○ Slow ○ Very Fast

Figure 1a. Ibigkas! Game Settings.



Score: 60

RAPID

Figure 1b. Ibigkas! Target word and choices

If the player's answer is correct, the background flashes green and the game proceeds to the next word until the time runs out (Figure 2a). If the player's answer is incorrect, the background flashes red and the player must choose another word until s/he gets the correct answer (Figure 2b).

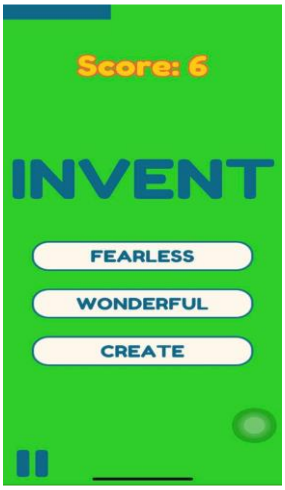


Figure 2a. Screen flashes green for correct word

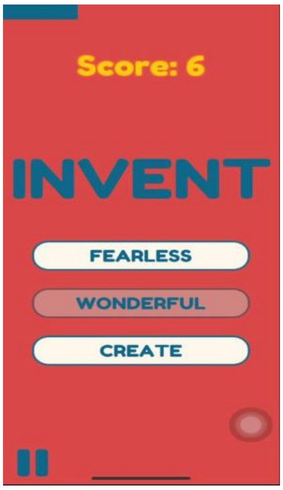


Figure 2b. Screen flashes red for incorrect word

In multiplayer mode, all the players must have their own mobile device and must be connected to the same network. One player must act as the host to choose the type of game. Each player would then receive three different words on their screen but one of them is also given the target word. The player with the target word must 'speak up' what target word is for all the other players to hear. The player with the correct word (rhyme, synonym, or antonym) shouts out the answer and tap it on his/her screen.

The Filipino version of *Ibigkas*! was later developed to help teach Filipino synonyms and antonyms (translated as '*magkasingkahulugan'* and '*magkasalungat'* in Filipino, respectively). The Filipino version follows the same mechanics as the English version.

3. Data Collection

The *Ibigka*s! games (English and Filipino versions) were tested on elementary students (grades 4, 5, and 6) from a rural public school and an urban public school in southern Philippines. The mother tongue of these students is Cebuano. Since there were no face-to-face classes due to the pandemic situation, a field staff was assigned to go to each of the participant's house to deliver the questionnaires and mobile device for testing while observing the required safety protocols. The researcher then communicated with the participants over the phone for orientation and instructions.

The participants first answered a demographics questionnaire to determine their level of access to mobile devices as well as their usage, attitude, and perceptions towards the English and Filipino languages. They were given statements like "I speak English at home" or "I speak Filipino at home" (with Cebuano translations) and they indicated their level of agreement using a five-point Likert scale (1=Strongly Disagree to 5=Strongly Agree). The participants then answered a pre-test on English and Filipino synonyms and antonyms. After which they were asked to play six *Ibigkas!* games: (1) English Rhymes, (2) English Synonyms, (3) English Antonyms, (4) Filipino Synonyms, and (5) Filipino Antonyms. The level of difficulty depends on the grade level: the 4th grade students played "medium" difficulty, the 5th graders played "hard" level, and the 6th graders played "very hard". All students played the games twice on the same difficulty setting but with different speeds, first with medium speed and then with fast speed setting. Interaction logs are automatically recorded on the device.

After playing the *Ibigkas!* games, the participants answered a post-test on English and Filipino synonyms and antonyms. Then they answered the Game-Based Learning (GBL) Engagement Metric (Chew, 2017) to determine how engaged the students were with the game. They were given statements like "When I was playing the games, I feel interested" and they indicated their level of agreement using a five-point Likert scale (1=Strongly Disagree to 5=Strongly Agree). Next, they were given the Intrinsic Motivation Inventory (IMI) (Ryan, 1982) questionnaire adapted for *Ibigkas!* with statements like "I tried very hard to answer correctly in the *Ibigkas!* games" and they indicated their level of agreement using a seven-point scale (1=Not at all true to 7=Very true).

4. Results and Discussion

4.1 Profile of Participants and Attitude towards English and Filipino

There were15 participants from each grade level (grades 4, 5, and 6) from urban and the same number from rural for a total of 90 participants. Out of the 90, 41 identified as female and 49 male. Only 17 from urban and 18 from rural have their own cellphone but more students from urban played mobile games including educational ones compared to those from rural (see Table 1).

Table 1. *Participants' Profile*

	Urban	Rural
Sex	Female = 18, Male= 27	Female = 23, Male = 22
Average age	11.13	10.38
Had their own mobile phone	17 (38%)	18 (40%)
Played mobile games	30 (67%)	21 (47%)
Played mobile educational games	23 (51%)	20 (44%)

4.1.1 Attitude towards English

The rural group has the smaller percentage of respondents who speak English at home (13.33%) and with their friends (13.33%) compared to Urban with 40.00% and 48.89%, respectively (see Table 2). When individual ratings were compared, a significant difference was found. In a survey conducted on 710 students from urban schools in National Capital Region, it was found that 35.41% speak some English at home (Rodrigo et al., 2019). The use of English at home or with friends may be more common in urban areas.

Also, more students from the urban group (80.00%) enjoy learning English than those from rural (51.11%). However, when asked if they enjoy reading in English, only 31.11% from urban while 57.78% from rural agreed to do so and the difference between groups was not significant.

Majority of the students from urban (57.78%) agreed that English is difficult to learn but only 8.89% disagreed, while the remaining 33.33% answered 'not sure'. The difference was significant when individual ratings were compared with rural group where 42.22% of the participants agreed and 42.22% disagreed. It is interesting to note that the urban group who use English more at home and with friends also found the language more difficult to learn whereas the rural group has the opposite observation. We looked at the relationship of the students' responses to the statement '*I find English difficult to learn*' and the average of their rating for using it '*at home*' and '*with friends*' and found a significant relationship, $r(88)=0.341$, $p<0.01$. Based on this, we can speculate that those who use it more are perhaps the ones more familiar to the level of difficulty of learning the language.

For both groups majority of the students expressed that they feel nervous when they need to speak English in class but all students expressed the desire to learn English and agreed that it is important to do so.

Table 2. *Attitude towards English: Urban vs Rural Comparison*

Questions	Urban		Rural		t-value	p-value
	mean	sd	mean	sd		
1. I speak English at home.	3.20	0.98	2.18	1.18	4.422	**<0.001**
2. I speak English with my friends.	3.31	1.11	2.22	1.09	4.632	**<0.001**
3. I enjoy learning English.	3.96	0.82	3.51	1.33	1.893	0.061
4. I enjoy reading in English.	3.44	1.27	3.36	1.59	0.289	0.774
5. I find English difficult to learn.	3.47	0.78	2.89	1.52	2.241	**0.027**
6. I feel nervous when I need to speak English in class.	3.44	1.04	3.42	1.54	0.079	0.938
7. I want to learn to speak and read in English.	3.91	0.78	3.96	1.28	-0.197	0.844
8. Learning English is important.	3.93	0.85	3.84	1.37	0.366	0.716

4.1.2 Attitude towards Filipino Language

Similar to the findings for the English language, rural has the smaller percentage of respondents who speak Filipino at home (13.33%) and with their friends (13.33%) and are significantly different when compared to urban with 42.22% and 55.56%, respectively (see Table 3). For both English and Filipino, results show that those from the urban areas speak the two languages more.

Most of the students from the urban group (80.00%) said they enjoy learning Filipino while 51.11% from the rural group said so but results show that students from both groups are not fond of reading in Filipino.

The students' ratings to the question about the difficulty of learning Filipino is very similar to how they feel towards the English language. Majority of the students from urban (57.78%) agreed that Filipino is difficult to learn while 20.00% disagreed, and the remaining 22.22% answered 'not sure'. The difference was significant when individual ratings were compared with rural where 42.22% of the respondents agreed and 42.22% disagreed, similar response to English. We also looked at the relationship of their response to the statement 'I find Filipino difficult to learn' and their usage (average of using it 'at home' and 'with friends') and found a significant relationship, *r*(88)=0.339, p=0.01. A similar speculation as with English could be made where those who use it more are perhaps the ones more familiar to the level of difficulty of learning Filipino.

Both groups also expressed feeling nervous when asked to speak Filipino in class but agreed that learning Filipino is important and expressed the desire to learn the language.

Table 3. *Attitude towards Filipino Language: Urban vs Rural Comparison*

Questions	Urban		Rural		t-value	p-value
	mean	sd	mean	sd		
1. I speak Filipino at home.	3.40	1.08	2.33	1.07	4.636	<0.001
2. I speak Filipino with my friends.	3.58	1.06	2.27	1.06	5.783	<0.001
3. I enjoy learning Filipino.	4.09	0.86	3.53	1.28	2.392	0.018
4. I enjoy reading in Filipino.	3.38	1.12	3.44	1.48	-0.238	0.812
5. I find Filipino difficult to learn.	3.62	0.80	3.04	1.37	2.424	0.017
6. I feel nervous when I need to speak Filipino in class.	3.60	1.06	3.51	1.42	0.332	0.740
7. I want to learn to speak and read in Filipino.	3.93	0.80	3.98	1.24	-0.200	0.842
8. Learning Filipino is important.	3.98	0.80	4.00	1.17	-0.104	0.918

Results show that the attitude towards English of the participants from each group is almost similar and to how they feel towards Filipino. The differences observed between the groups are also the same for both languages.

4.2 Prior Knowledge

Despite of the differences in their attitude towards the languages, no significant difference was found on the pre-test performance between the groups for all grade levels. It is also interesting to note that no average percentage is higher than 71% which is relatively low. The overall averages are even lower: for the urban group it's 45% for Ibigkas! (English) and 57% for Ibigkas! Filipino while for the rural group, the averages are 45% and 56%, respectively. However, given the 15% to 25% standard deviations, we can say that the scores are varied. The fact that, overall, the majority of the participants do not speak neither English nor Filipino at home or with friends and are not fond of reading in both languages may have contributed to these scores.

4.3 In-Game Performance

To account for in-game performance, we look at the student's attempts per given word. If the student answers correctly on the first attempt, 1 point is awarded; a half point on the second attempt, and zero for the third attempt since the correct answer must have been revealed at this point. The percentage was obtained by dividing the computed score by the total number of words given. All students played two levels of speed for the level of difficulty assigned in their grade level. For example, a grade 4 students were assigned medium difficulty for rhymes, synonyms, and antonyms and they have to play the games first with medium-level speed then fast-level speed. We then computed the average of the two speeds for each game per student to measure their performance per game type.

4.3.1 Comparison of In-game Performance between Urban and Rural

There was no significant difference on in-game performance between urban and rural for the *Ibigkas!* English games in all levels. For *Ibigkas!* Filipino, no significant difference was found for all games although there were differences in their attitude towards Filipino. Since there was also no significant difference found in the pre-test, this result is not totally surprising. It is also interesting to note that all average percentages for all grade levels are lower than 71% for urban and lower than 66% for rural.

The overall group averages are even lower. For the urban group it is 52% for *Ibigkas!* (English) and 57% for Ibigkas! Filipino while for the rural group are 59% and 55%, respectively.

The in-game performance of the students also reflect their pre-test performance. Considering the overall attitude of the participants towards English and Filipino as well as their usage, these numbers are to be expected especially that it was their first time to play the game.

4.3.2 Analysis of Attempts

Due to relatively low in-game performance, we look at the attempt level of the logs to understand how the students identified the correct answer to the given words. Since the players were given 3 options, they have a total of 2 chances to choose the correct word before the game would reveal the correct word. Specifically, we look at the incidence of correct-at-first-attempt and correct-at-second-attempt. We found that for all the games across grade levels and groups, majority of the correct answers were on the second attempt (see Table 4). This means that their first choice was revealed by the game to be incorrect by providing a visual cue of a flashing red background. On the second attempt, they are now left with two choices with a higher probability of getter the correct answer. By looking at the probability in a straightforward manner, we could say that during the first attempt, the probability of choosing the correct word from the three choices is 33.33%. During the second attempt with only two options left, we could say the probability increased to 50%. However, this approach does not totally explain the higher incidence of correct answers in the second attempt across all games and groups.

A similar concept can be found in the Monty Hall Problem where the player is on a game show and is given the choice of three doors. Behind one door is a car, and behind the other two are goats. Say, the player chose door 1, and the host, who knows where the car is, opens door 3, behind which is a goat (Selvin, 1975). According to the solution presented by Vos Savant (1990), the probability of choosing the correct door between the two remaining doors after the host revealed a wrong door is not 50% but actually 33.33% if you stay with your choice and 66.66% if you switch to the other remaining door.

Considering that the *Ibigkas!* gameplay and the Monty Hall mechanics have differences and that the probabilities in Monty Hall are heavily dependent on its context, we cannot fully assume that the probabilities apply to the *Ibigkas!* observations but this gives us an idea of looking at the *Ibigkas!* scenario from a different perspective. Given that there are 3 choices and the students are afforded 2 chances (2 out of 3) before the game reveals the correct answer, we can say that the players actually have a total of 66.66% probability to get the correct word. Hence, it is more likely that the students would get the correct answer by the second attempt.

Table 4. *Incidence of correct answer on the 1st and 2nd attempts*

Groups	Urban			Rural		
Games	Correct on 1st attempt	Correct on 2nd attempt	Game revealed the answer	Correct on 1st attempt	Correct on 2nd attempt	Game revealed the answer
English						
G4 Rhymes	31.72%	54.63%	13.66%	32.59%	58.27%	9.14%
G4 Synonyms	14.36%	70.79%	14.85%	16.60%	73.29%	10.60%
G4 Antonyms	24.67%	64.32%	11.01%	26.55%	61.28%	12.18%

G5 Rhymes	39.75%	52.17%	8.07%	46.24%	46.24%	7.52%
G5 Synonyms	13.38%	69.43%	17.20%	23.76%	66.30%	9.94%
G5 Antonyms	17.07%	65.24%	17.68%	18.31%	68.92%	12.77%
G6 Rhymes	33.14%	52.07%	14.79%	26.91%	64.00%	9.09%
G6 Synonyms	20.51%	59.62%	19.87%	13.76%	75.05%	11.18%
G6 Antonyms	12.67%	72.00%	15.33%	20.10%	69.28%	10.62%
Filipino						
G4 Synonyms	23.23%	58.59%	18.18%	12.38%	77.18%	10.44%
G4 Antonyms	28.57%	60.00%	11.43%	17.03%	71.82%	11.15%
G5 Synonyms	35.07%	57.46%	7.46%	18.21%	68.07%	13.72%
G5 Antonyms	40.43%	51.06%	8.51%	22.84%	68.07%	9.09%
G6 Synonyms	25.78%	63.28%	10.94%	14.77%	76.58%	8.65%
G6 Antonyms	23.90%	62.89%	13.21%	17.58%	70.97%	11.44%

Moreover, based on the percentages of correct answers on the first attempt, it was observed that students for all groups and grade levels performed best on rhymes, followed by antonyms, then synonyms for *Ibigkas!* English. For *Ibigkas!* Filipino, majority tend to perform better on antonyms than in synonyms. This observation is somehow opposite from the prediction of some teachers from a predominantly Tagalog-speaking region that students would find antonyms more difficult than rhymes and synonyms (Rodrigo et al., 2019).

4.4 Game-Based Learning Engagement and Intrinsic Motivation Inventory

Across all groups and grade-level, the students generally gave a positive feedback towards the games. Results from the game-based learning engagement questionnaire reveal that they carefully followed the instructions (3.9/5.0) and that they tried their best to identify the correct answer (3.9/5.0). They also agreed that when playing the game, they are able to use and practice what they have learned in class (4.2/5.0) and that the game helped them widen and improve their English and Filipino vocabulary (3.8/5.0). They though that game was interesting (3.8/5.0) that they look forward to completing the tasks (4.6/5.0), and that the games have enough difficulty to challenge them (4.1/5.0).

The same positive response is reflected in the Intrinsic Motivation questionnaire. The students enjoyed playing the games (5.4/7.0), described them as interesting (5.8/7.0), and thought they were fun to play (5.9/7.0). They said that they tried their best (5.9/7.0) because it was important for them to do well (5.3/7.0) and they believed they did pretty good (5.7/7.0).

Engagement features were derived from the GBLE responses and motivation features from the IMI responses as described in (Moreno et al., 2019): behavior engagement (i.e. being attentive and trying their best to identify the correct word), cognitive engagement (i.e. using and applying what they have learned in class while playing the game, asking questions when they didn’t know what to do, and thinking that the game has enough difficulty to challenge them), emotion engagement (i.e. being interested while playing and looking forward to finish the game), enjoyment (i.e. enjoying, having fun and finding the game interesting and not boring), effort (i.e. trying hard to answer correctly and thinking that it was important for them to do well), and perceived competence (i.e. believing that they played *Ibigkas!* pretty well and feeling satisfied with their performance). In the comparison between the two groups, no significant difference was found in terms of the GBLE features. For the IMI features, participants from the urban group had significantly higher self-report ratings for enjoyment (*M*=5.69, *SD*=1.01) than the rural group (*M*=4.93, *SD*=1.83, *t*(88)=2.40, *p*=0.02) and the same for effort where the urban group had higher ratings (*M*=5.92, *SD*=0.97) than the rural group (*M*=5.29, *SD*=1.48, *t*(88)=2.37, *p*=0.02).

4.5 Learning Gains

No significant difference was found on the learning gains between the groups except for 6th grades students from urban who had significantly higher gains on Filipino Synonyms (the highest among all learning gains) compared to the rural 6th grade students. Although there a number of negative gains, the game yielded positive gains up to 25% considering that it is the first time for the participants to be exposed to and play the game and were only able to play one round for each type, level, and speed assigned to them.

The positive GBL Engagement and IMI feedback as well as some learning gains can be considered a promising step on the use of mobile games to improve English and Filipino vocabulary among learners who are non-native speakers of the languages.

5. Summary and Conclusion

Urban students tend to use English and Filipino more than rural students but they share the same sentiments as to the importance of learning both languages. We also found that students who used the languages more, are the ones who thought the languages were difficult to learn. We speculate that due to their exposure, they are more acquainted with the components of language learning. Both groups said that they get nervous when they have to speak the languages in class but they also agreed that learning English and Filipino can be enjoyable. Moreover, the fact that majority of the participants do not speak neither English nor Filipino at home or with friends and that they are not fond of reading in the said languages may have contributed to their low pre-test scores.

Their low in-game scores also reflect the participants' low prior knowledge and the overall usage and attitude towards English and Filipino. Upon further investigation of their attempts, it was found that for all the games across grade levels and groups, majority of the correct answers were on the second attempt. This could be attributed to the game mechanics that allows the students a total of 2 attempts given the 3 choices which gives them 66.66% probability of getting the correct answer on the second attempt. Also, the particular mechanic of the game that allows students to try again until the correct answer is revealed may have contributed to some of the positive learning gains considering that it was their first time to play the game.

The findings showed that: (1) urban-based students speak English and Filipino more at home and with their friends, they enjoy learning the both languages more, but they also find them difficult to learn more than those from rural areas (2) there is promising potential on the use of the *Ibigkas!* mobile games in improving the students' Filipino and English vocabulary as shown by the students GBL engagement and IMI responses as well as some positive learning gains from their first time playing the games, and (3) overall, no significant difference was found on in-game performance between urban and rural for all the Filipino games across all levels and most of the English games except between 5th Grade English Rhymes and 5th Grade English Synonyms.

In conclusion, results are consistent with the first hypothesis that urban-based students accept and use the languages more than rural-based students. On the other hand, it failed to prove the second hypothesis that urban-based students have better language skills than rural students as generally, no significant difference was found between their English and Filipino pre-tests, in-game performance, and post-tests.

Acknowledgements

We thank the Ateneo de Manila University Research Council for the post-doctoral grant of MMPTF that made this study possible; the Ateneo Laboratory for the Learning Sciences of Ateneo de Manila University for the *Ibigkas!* games and for the mobiles phones used in the field test; the students of the public schools in Bukidnon, Philippines who participated in this study; and Kent Levi A. Bonifacio and Gladys S. Ayunar of Central Mindanao University for their help in the data collection.

References

Ball, J. (2010). Enhancing learning of children from diverse language backgrounds: Mother tongue-based bilingual or multilingual education in early childhood and early primary school years. Early Childhood Development Intercultural Partnerships, University of Victoria.

Barnachea, A. A. (2013). Philippines' Public School Curriculum Model. Retrieved from https://www.slideshare.net/TeacherAdora/curriculum-models-philippines-curriculum-models.

Barron, S. (2012). Why language matters for the Millennium Development Goals (MDG). Bangkok: UNESCO.

Chew, B. S. (2017). An efficient framework for game-based learning activity. *2017 IEEE 6th International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 147–150. IEEE.

Dekker, D., & Young, C. (2005). Bridging the gap: The development of appropriate educational strategies for minority language communities in the Philippines. *Current Issues in Language Planning, 6*(2), 182–199.

Giannakas, F., Kambourakis, G., Papasalouros, A., & Gritzalis, S. (2018). A critical review of 13 years of mobile game-based learning. *Educational Technology Research and Development, 66*(2), 341–384.

Huang, C.-J., Chen, H.-X., & Chen, C.-H. (2009). Developing argumentation processing agents for computer-supported collaborative learning. *Expert Systems with Applications, 36*(2), 2615–2624.

Jubilado, R. C. (2004). Philippine linguistics, Filipino language, and the Filipino nation. *Jati Journal of Southeast Asian Studies, 9*, 43–54.

Kukulska-Hulme, A., Lee, H., & Norris, L. (2017). Mobile learning revolution: Implications for language pedagogy. In C. A. Chapelle & S. Sauro (Eds.), *The handbook of technology and second language teaching and learning* (pp. 217–233). Hoboken: John Wiley & Sons. Retrieved from http://dx.doi.org/10.1002/9781118914069.ch15

McFarland, C. D. (2008). Linguistic diversity and English in the Philippines. *Philippine English: Linguistic and Literary Perspectives, 1*, 131.

Metila, R. A., Pradilla, L. A. S., & Williams, A. B. (2016). The challenge of implementing mother tongue education in linguistically diverse contexts: The case of the Philippines. *The Asia-Pacific Education Researcher, 25*(5), 781–789.

Miangah, T. M., & Nezarat, A. (2012). Mobile-assisted language learning. *International Journal of Distributed and Parallel Systems, 3*(1), 309.

Moreno, M., Manahan, D., Fernandez, M., Banawan, M., Beraquit, J., Caparos, M., … Rodrigo, M. M. T. (2019). Development and Testing of a Mobile Game for English Proficiency Among Filipino Learners. *Proceedings of the 27th International Conference on Computers in Education*. Presented at the 27th International Conference on Computers in Education, Taiwan.

Mouza, C., & Barrett-Greenly, T. (2015). Bridging the app gap: An examination of a professional development initiative on mobile learning in urban schools. *Computers & Education, 88*, 1–14.

Ozdamli, F., & Cavus, N. (2011). Basic elements and characteristics of mobile learning. *Procedia-Social and Behavioral Sciences, 28*, 937–942.

Philippine Department of Education. (2019). *PISA 2018 National Report of the Philippines* [Department of Education Complex, Meralco Avenue, Pasig City Philippines]. Retrieved from https://www.deped.gov.ph/wp-content/uploads/2019/12/PISA-2018-Philippine-National-Report.pdf

Philippines Department of Education. (2009). Institutionalizing mother tongue-based multilingual education (MLE). *Philippines Department of Education*. Retrieved from http://www. deped.gov.ph/orders/do-74-s-2009

Rodrigo, M. M. T., Ocumpaugh, J., Diy, W. D., Moreno, M., De Santos, M., Cargo, N., … Beraquit, J. I. (2019). Ibigkas!: The Iterative Development of a Mobile Collaborative Game for Building Phonemic Awareness and Vocabulary. *Computer-Based Learning in Context, 1*(1), 28–42.

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology, 43*, 450–461.

Selvin, S. (1975). A problem in probability (letter to the editor).

Sykes, J. M. (2017). Technologies for teaching and learning intercultural competence and interlanguage pragmatics. *The Handbook of Technology and Second Language Teaching and Learning, 118*, 133.

Tupas, R., & Lorente, B. P. (2014). A ‘new’politics of language in the Philippines: Bilingual education and the new challenge of the mother tongues. In Language, Education and Nation-building (pp. 165–180). Springer.

Vos Savant, M. (1990). Game Show Problem. Parade. Retrieved from https://web.archive.org/web/20130121183432/http://marilynvossavant.com/game-show-problem/

World Bank. (2005). *In their own language: Education for all*. The World Bank. Retrieved from The World Bank website: https://openknowledge. worldbank.org/handle/10986/1033.

Analytics of Open-Book Exams with Interaction Traces in a Humanities Course

Rwitajit MAJUMDAR[a*], Geetha BAKILAPADAVU[b], Jiayu LI[c]
Mei-Rong Alice CHEN[d], Brendan FLANAGAN[a] & Hiroaki OGATA[a]
[a]Academic Center for Computing and Media Studies, Kyoto University, Japan
[b]Department of Humanities and Social Science, BITS Pilani K K Birla Goa Campus, India
[c]Graduate School of Informatics, Kyoto University, Japan
[d]National Taiwan University of Science and Technology, Taiwan
*dr.rwito@gmail.com

Abstract: Open book exams (OBE) have been a mandated part of each course structure at some universities. Also during the COVID19 emergency remote teaching situation, OBE would be an option for many instructors over a proctored examination. In this study we investigate a *Critical Analysis of Literature and Cinema* course which had offered open book exam components for more than 11 years in a face-to-face classroom mode. However, this time the OBE was conducted online using BookRoll, a learning analytics enhanced eBook platform. 89 Students accessed *Hayavadana,* an Indian play uploaded on BookRoll during the exam. They attempted a critical reading task to identify performative elements and cultural references in the text by highlighting them with yellow and red markers respectively and writing a reflective memo about the identified items in BookRoll. We analyzed learner's interaction logs gathered in the learning record store linked to BookRoll during the OBE and investigated the relations between their critical reading behaviors to the OBE achievement. Further, selecting two distinct achievement groups we conducted process mining to identify distinct reading behaviors of the high and low performers and give examples of their generated reflective memos. This study aims to initiate further discussion related to the application of learning analytics in humanities courses and probed into the behaviors of the students during the OBE.

Keywords: Learning analytics, open book exam, humanities course, BookRoll, critical reading activity, process mining, Hayavadana

1. Background and Motivation

From the humanities education standpoint, developing critical reading skills is crucial. Critical reading is an active, in-depth reading of a text that calls forth a deeper engagement with the text. Such an activity requires cognitive tasks such as comprehending, analyzing, evaluating, interpreting and synthesizing. Critical reading enables the reader to read not only the explicit meanings but the layered and the implicit meanings as well. One of the essential values of Humanities is identified as critical thinking (Holm et.al., 2015). Especially so in the case of courses that deal with cultural texts including narrative arts. While understanding subjective experiences embedded in the texts is important, deciphering the layered meanings in them is equally significant. In that context, a pertinent way to assess the student is not only to ask them to remember and reflect on the text but provide an open book exam (OBE) where the text is provided during the process of reflection itself. Analysis of students' behavior during OBE to understand their critical reading behaviors given a focused evaluative activity has the potential to thereafter design learning scaffolds. However limited scholarship investigates this process. It would not only require a meaningful reading task but also some technical affordance to trace the learners' behaviors. BookRoll, an ebook reader linked in the learning and evidence analytics framework fitted as an appropriate technology choice. The current study looks at an actual implementation of an OBE in the BookRoll environment and collects interaction traces of students who attempted the task. In such an authentic learning assessment setting, we investigate the following two research questions:

1. What are the relations between the reading behaviors and achievements of the learners in an OBE with critical reading tasks?

2. What are the differences between the high and low performing students during an OBE with a critical reading task to identify cultural references and performative elements based on their interaction patterns and outputs during attempting the task?

The article is organized in the following sections. Section 2 looks at the related works and provides the foundation of the study. Section 3 illustrates the course context of the OBE and the research methods. Section 4 presents the results of the analysis. Section 5 ends with a discussion and conclusion of the study.

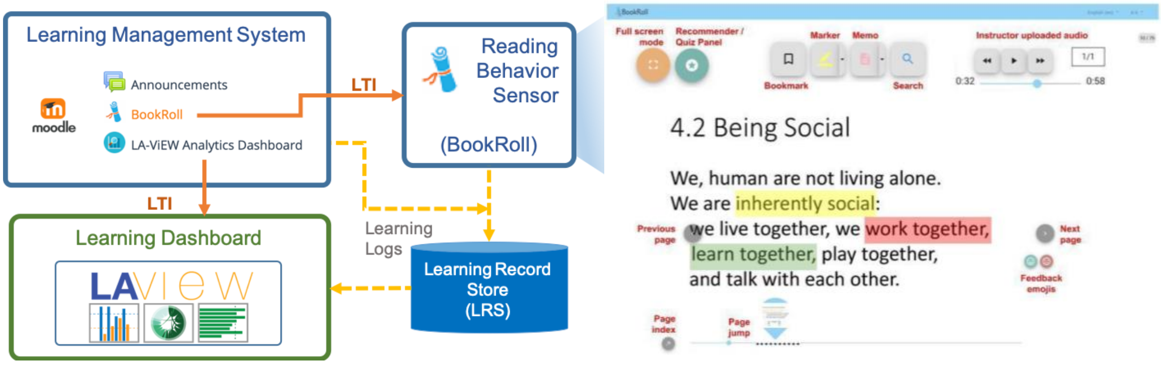
2. Related Work and Foundation of the Study

2.1 Open Book Exams (OBE)

OBE, in a broad sense, allows the students to open either or all of their study material, text book, class notes and other reference material. Chan (2009) points out OB exams are less demanding on memory as it is no longer necessary for students to cram a lot of facts, figures and numbers, provides a chance for students to acquire the knowledge during the preparation process itself, enhances information retrieval skills of students and enhances the comprehension and synthesizing skills of students. OBE are often designed to call forth higher cognitive levels and to promote study and teaching methods that would improve understanding (Eilertsen & Valdermo 2000). Earlier researchers compared open-book and closed-book exams (Theophilides & Koutselini 2000; Block, 2012). They found that students preparing for an open-book examination tend to consult various sources and interrelate the information acquired and while taking the exam, they work creatively and probe deeply into the knowledge gained. In one study, OBE was conducted online and the effect of training on the OBE was investigated; the effect of training was found to have positive outcomes where the ones who were trained scored higher (Rakes, 2008). While as a concept OB exams are not new, very few institutions have been using them on a regular basis. The university where the current study was conducted has been one of the very few institutions to implement OBE across disciplines for more than 35 years (Improbable Achievements, 1990). These exams provide a range of flexibilities wherein these exams can be fully or partially open book. The situation that rose due to the Covid 19 pandemic made many universities explore employing OBE. In an online, remote exam scenario OBE is a favorable way of administering assessment as it can minimize use of unfair means while attempting an exam. In one such study during the emergency remote teaching due to COVID-19, researchers investigated the effect of OBE (Ashri & Sahoo, 2021).

2.2 Learning Evidence Analytics Framework (LEAF)

Learning Evidence Analytics Framework (LEAF) is an overarching technology framework to collect evidence of learning and teaching from the logs generated in a technology-enhanced learning environment (Ogata et al. 2018). In this instantiation of the framework, the instructor coordinated the course on the university's Learning Management System (LMS), Moodle. BookRoll, an e-book reader, was linked to LMS via Learning Tools Interoperability (LTI) protocol and used to upload the reading contents like lecture slides, reference articles and reading assignments in PDF format for students to access. Tools like BookRoll can be considered as a learning behavior sensor as it can log student's reading and annotation interactions in a Learning Record Store (LRS) as standard eXperience API (xAPI) statements. Figure 1 presents the technical architecture based on LEAF that is used in our study and the user's reading interface in BookRoll which supports annotation functions such as highlighting with different colors, adding memos and bookmarks in the content. As long as there is an internet connection, students can read their books anytime from a web browser on their personal computer or smartphones. Student's reading activity log from the LRS is then provided to the dashboard database and visualized for both the instructors and students appropriately.



4.2 Being Social

We, human are not living alone.
We are inherently social:
we live together, we work together,
learn together, play together,
and talk with each other.

Figure 1. Learning and Evidence Analytics Framework (LEAF) and BookRoll interface.

3. Study Context and Method

3.1 Context and Participants

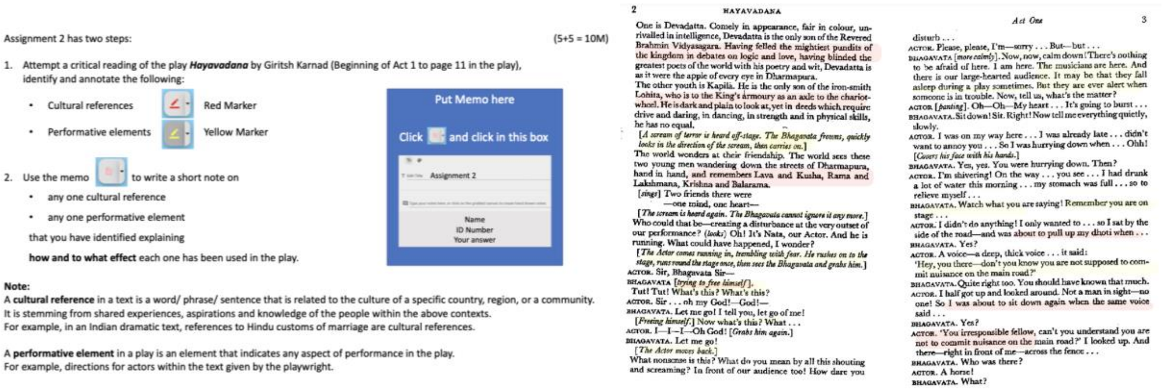
This particular study furthers the ongoing research on reading analytics of learners in a Humanities education context at the undergraduate level (Majumdar et.al. 2020, 2021). Critical Analysis of Literature and Cinema (CALC) is an elective course offered by the Humanities and Social Sciences Department in a private University in India. The course CALC encompasses the following objectives; inculcate in students a critical insight required to interpret a work of literature and cinema, enable the students to perceive the subtle nuances of such works and to develop critical judgment, and to introduce different forms, terminologies and trends prevalent in such artistic ventures to enable them to place a work of art in the proper context. Reading tasks, film viewing tasks, and critical analysis activities are integral parts of this course. Within this context, it is mandatory to keep a portion of the evaluative components as open book. The class was scheduled for 3 hours each week, split across three sessions. Students met for a total of fifteen weeks. In addition to these classroom interactions, students were given take-home readings and film viewings. The classes had to be conducted in an online mode for the entire semester due to the Pandemic Covid 19 situation. The exams and other assignments were conducted using the LMS. In this particular case one of the evaluative components was carried out using Bookroll.

We followed a purposive sampling and selected the students enrolled in the course (n=89, 77 males, 12 females) as participants. They were pursuing their undergraduate program in engineering and sciences in the university. The class included students in the age group of 19 to 23 years and were enrolled in their second, third or fourth year of study in the university. At the time of the research, they had been introduced to approximately 0 to 3 humanities courses as electives. For the open-book exam the students could access BookRoll from their course Moodle. The details of the exam and instructions are as follows.

An overall phenomenographic research approach (Jan Larsson & Inger Holmström, 2007) guided the research questions to focus on a single activity undertaken by the students enrolled in the course, the OBE in this case. The team of researchers including the course instructor then interpreted the different approaches that emerged from the interaction logs during that episode of the exam.

3.2 Open-Book Exam: Instructions and Example of Critical Reading Task Given

The instructor chose an Indian play titled Hayavadana (Karnad, 1972), originally written in Kannada and then translated into English by the playwright himself. The content was uploaded on BookRoll and the students were given access to the content 3 days before the actual open book evaluative component. The designed task involved students going through the first act of the play to first identify and highlight cultural references (red highlight) and performative elements (yellow highlight) and then to write a memo each on both these elements in the text designated. The activity was designed around these two factors as the play is deeply steeped in the cultural milieu of traditional Indian theatre. Also, Hayavadana being a densely multi-layered text, the instructor had identified these two tasks to be significant in the context of deciphering layered meanings of the text in a critical reading scenario. Instructions for the one hour OBE were given to the students posted on the LMS followed by announcing them to do it during the online synchronous class (Fig 2a.). An example of the highlighted cultural reference and the performative element is also shown below (Fig 2b.). We have selected 2 of the pages which the students have spent most of the time (see section 4.1)



Assignment 2 has two steps: (5+5 = 10M)

1. Attempt a critical reading of the play ***Hayavadana*** by Giritsh Karnad (Beginning of Act 1 to page 11 in the play), identify and annotate the following:
   - Cultural references Red Marker
   - Performative elements Yellow Marker
2. Use the memo to write a short note on
   - any one cultural reference
   - any one performative element

   that you have identified explaining

   how and to what effect each one has been used in the play.

Note:
A **cultural reference** in a text is a word/ phrase/ sentence that is related to the culture of a specific country, region, or a community.
It is stemming from shared experiences, aspirations and knowledge of the people within the above contexts.
For example, in an Indian dramatic text, references to Hindu customs of marriage are cultural references.

A **performative element** in a play is an element that indicates any aspect of performance in the play.
For example, directions for actors within the text given by the playwright.

2 HAYAVADANA

One is Devadatta. Comely in appearance, fair in colour, unrivalled in intelligence, Devadatta is the only son of the Revered Brahmin Vidyasagara. Having felled the mightiest pundits of the kingdom in debates on logic and love, having blinded the greatest poets of the world with his poetry and wit, Devadatta is as it were the apple of every eye in Dharmapura.
The other youth is Kapila. He is the only son of the iron-smith Lohita, who is to the King's armoury as an axle to the chariot-wheel. He is dark and plain to look at, yet in deeds which require drive and daring, in dancing, in strength and in physical skills, he has no equal.

[*A scream of terror is heard off-stage. The Bhagavata frowns, quickly looks in the direction of the scream, then carries on.*]

The world wonders at their friendship. The world sees these two young men wandering down the streets of Dharmapura, hand in hand, and remembers Lava and Kusha, Rama and Lakshmana, Krishna and Balarama.

[*sings*] Two friends there were
—one mind, one heart—

[*The scream is heard again. The Bhagavata cannot ignore it any more.*]

Who could that be—creating a disturbance at the very outset of our performance? (*looks*) Oh! It's Nata, our Actor. And he is running. What could have happened, I wonder?

[*The Actor comes running in, trembling with fear. He rushes on to the stage, runs round the stage once, then sees the Bhagavata and grabs him.*]

ACTOR. Sir, Bhagavata Sir—
BHAGAVATA [*trying to free himself*].
Tut! Tut! What's this? What's this?
ACTOR. Sir . . . oh my God!—God!—
BHAGAVATA. Let me go! I tell you, let go of me!
[*Freeing himself.*] Now what's this? What . . .
ACTOR. I—I—I—Oh God! [*Grabs him again.*]
BHAGAVATA. Let me go!
[*The Actor moves back.*]
What nonsense is this? What do you mean by all this shouting and screaming? In front of our audience too! How dare you

Act One 3

disturb . . .
ACTOR. Please, please, I'm—sorry . . . But—but . . .
BHAGAVATA [*more calmly*]. Now, now, calm down! There's nothing to be afraid of here. I am here. The musicians are here. And there is our large-hearted audience. It may be that they fall asleep during a play sometimes. But they are ever alert when someone is in trouble. Now, tell us, what's the matter?
ACTOR [*panting*]. Oh—Oh—My heart . . . It's going to burst . . .
BHAGAVATA. Sit down! Sit. Right! Now tell me everything quietly, slowly.
ACTOR. I was on my way here . . . I was already late . . . didn't want to annoy you . . . So I was hurrying down when . . . Ohh! [*Covers his face with his hands.*]
BHAGAVATA. Yes, yes. You were hurrying down. Then?
ACTOR. I'm shivering! On the way . . . you see . . . I had drunk a lot of water this morning . . . my stomach was full . . . so to relieve myself . . .
BHAGAVATA. Watch what you are saying! Remember you are on stage . . .
ACTOR. I didn't do anything! I only wanted to . . . so I sat by the side of the road—and was about to pull up my dhoti when . . .
BHAGAVATA. Yes?
ACTOR. A voice—a deep, thick voice . . . it said:
'Hey, you there—don't you know you are not supposed to commit nuisance on the main road?'
BHAGAVATA. Quite right too. You should have known that much.
ACTOR. I half got up and looked around. Not a man in sight—no one! So I was about to sit down again when the same voice said . . .
BHAGAVATA. Yes?
ACTOR. 'You irresponsible fellow, can't you understand you are not to commit nuisance on the main road?' I looked up. And there—right in front of me—across the fence . . .
BHAGAVATA. Who was there?
ACTOR. A horse!
BHAGAVATA. What?

Figure 2a. Instructions *2b*. Examples of Cultural References and Performative Elements in a Page.

3.3 Data Collection and Analysis

The data extracted for this study included the reading logs in BookRoll during the open book exam period. Of the total 89 students enrolled in this course, 84 accessed BookRoll to answer the open book exam. 15071 logs were captured during the activity. For this study, we considered the 14838 action logs of opening, navigating through its pages and annotating its content. The instructor evaluated and scored the answers out of 10 marks, 1 mark each for marking and 4 marks each for the memo. The interactions and score distribution are presented in Table 1.

Table 1. *Distribution of the Students' Action and Score during the Open-Book Exam*

	Time (min)	Completion (%)	Events (counts)	Red Markers	Yellow Markers	Memo	Bookmark	Score
Mean	59.798	42.45	188.845	29.2	44.9	2.1	0.2	7.43
Std. Dev	22.752	19.28	59.504	14.1	26.6	1.0	0.5	1.52
Minimum	28	21	79	3	3	0	0	0
Maximum	135	95	344	69	118	6	2	9.5

In our earlier work we characterised critical reading behavior of the learners (Majumdar et al. 2021). In the current study with a new batch of students the critical reading based activity was conducted as an OBE and had a time constraint. Hence we aim to investigate more about the process of attempting the highlight and memo based tasks. To answer RQ1 the interaction log was collected and processed to get the interaction counts of each annotation action and its correlations to the outcome score as computed by JASP (JASP Team, 2019). Then DISCO (Fluxicon, 2021) was used to find the prominent interaction process for the cohort while they attempted the OBE. To answer RQ2, and understand the difference between students who were at different performance levels, we separated the group into high and low groups based on the score provided after the instructor evaluated all the highlights and the reflective notes. Based on the performance distribution of the whole class, the students who scored 8 (out of 10) or above were considered as high performers (n=33) and those who scored 7 or below were low performers (n=26). The choice to leave the mid performers (n=25) was to consider a clear distinction of the performance groups.

4. Results and Interpretations

4.1 Correlation between Student's Reading Behaviors and Achievement in OBE

In the given dataset, the correlation between the student reading behaviors in navigating, annotating, editing annotations and the achievement as measured by the OBE score is analyzed and presented in Table 2. Count of number of interactions and the total duration of each of Navigation behavior (actions consisting of NEXT, PREV, PAGE_JUMP and BOOKMARK_JUMP), Annotation behavior (actions consisting of ADD_MARKER, ADD_MEMO and ADD_BOOKMARK) and Editing behavior (actions consisting of CHANGE MEMO, DELETE_MEMO or DLETE_MARKER) is computed. Considering the significance level at p=0,05, the achievement score was positively correlated to the duration of editing behavior. The count of the editing behavior was positively related to its duration and also to the count of navigation. The count and the duration of the annotation behavior are negatively correlated, whereas the annotation count is positively correlated to the time spent on navigation. Navigation behavior's count and duration are also negatively correlated.

Table 2. *Correlation between Interaction Behaviors and Achievement in OBE (n=84 students)*

		Count of interactions (_c)			Duration of interactions (_d)		
	OBE Score	Navigation	Annotation	Editing	Navigation	Annotation	Editing
OBE Score	—						
Navigation_c	0.035	—					
Annotation_c	0.061	-0.111	—				
Editing_c	0.211	0.341**	0.105	—			
Navigation_d	0.038	-0.248*	0.265*	0.002	—		
Annotation_d	0.019	0.104	-0.392***	-0.08	-0.073	—	
Editing_d	0.23*	-0.03	0.055	0.248*	0.027	-0.07	—

$^*p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$

4.2 Interaction Patterns during the Open-Book Examination

To answer the OBE, three subtasks for the students would be i. reading the allotted pages for understanding the text and its nuances, ii. identifying cultural references and performative elements in those pages as markers and iii. writing one memo each on both the marked elements. It typically would involve comprehension, synthesis and reflexive tasks. Through process mining from the interaction logs as shown in Fig 4, the process of the interaction behaviors emerged. The process mining overview has each interaction logged as a state, represented as a node in the graph and the sequence (transition of one action to another) as the edge of the graph. The information in the node also provides the number of individuals that did the specific action. For instance 84 students opened the text. The information in the edge presents the average time between the transition to the next action and how many students had the specific transition pattern. For instance, after opening the text and reading the content for an average of 14.3seconds, 9 students used the slider to navigate to another page. The edge width and the color intensity is proportional to the mean duration.

Considering the states from the diagram for all the students' (n=84) interactions during OBE, it shows that while all the participants used NEXT and PREVIOUS action to navigate, few of them used JUMPs either by using slider (n=30, 35%) or by using the BookMark that they added (n=4, 5%). For annotations all of them did yellow and red markers but only n=81, 96% of them added memos. As for modifying annotations, around 60% deleted the yellow (n=51) or Red markers (n=49) at some point of time. As for memos 70 (83%) students changed them and 23(27%) students deleted.

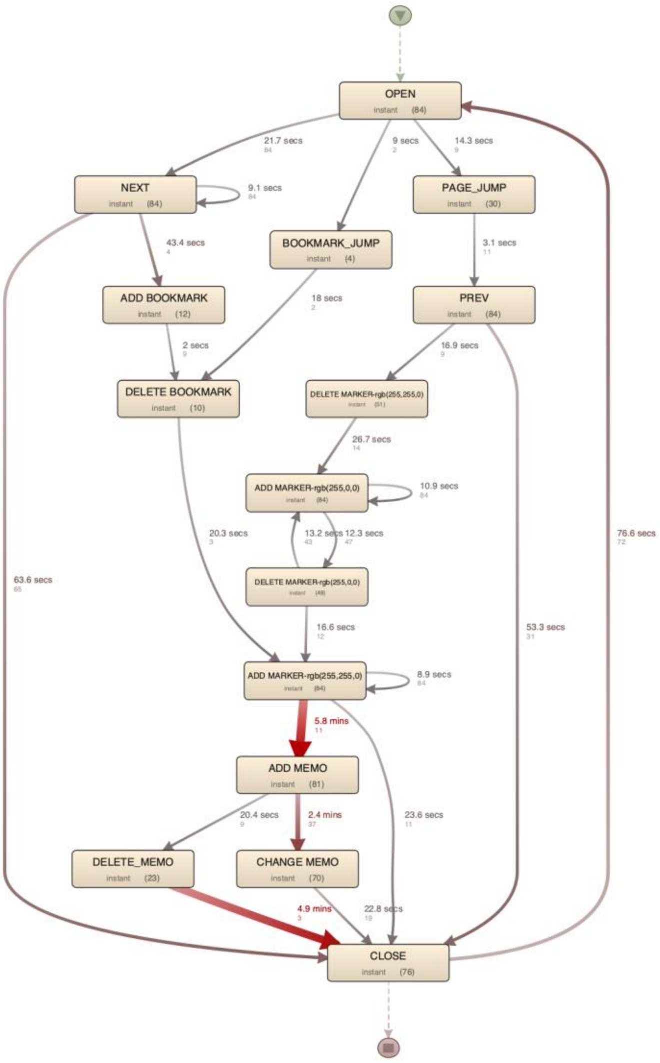


Figure 3. Overall Interaction Sequence during OBE (n=84).

As for the information regarding the average time spent on each interaction, first we checked the actions which were repeated. Consecutive clicking NEXT, adding Markers (both yellow and red) are seen for all the participants. On average consecutive NEXTs were clicked after 9.1 seconds. Adding red markers and yellow markers were after 10.9 and 8.9 seconds respectively. Adding and deleting red markers consequently were seen in more than 50% students with around 12.75 seconds as mean interval. Adding the memo was 5.8 mins for 13% students. 44% of students changed the memo after 2.4 minutes.

4.3 Differences in Interaction Patterns and Outcomes of High and Low Performers.

To answer RQ2, the interaction process of both the high and low groups were mined. The criteria of selecting an activity (interaction state) node in the process was set to 50%, that is at least 50% of the group members had it. The top 30% prominent paths (links) were selected. We present the interaction behaviors of the two performance groups and examples of the reflective memo submitted.

4.3.1 Low Performers Interaction Patterns and Artefact Created

For the low performing group the interaction sequence had three initiating actions related to navigating or adding red or yellow markers (see Fig 5). The adding memo action had an average duration of 3.8 minutes. It was also seen that 55.7% of all the actions happened in pages 9 to 11 of the content.

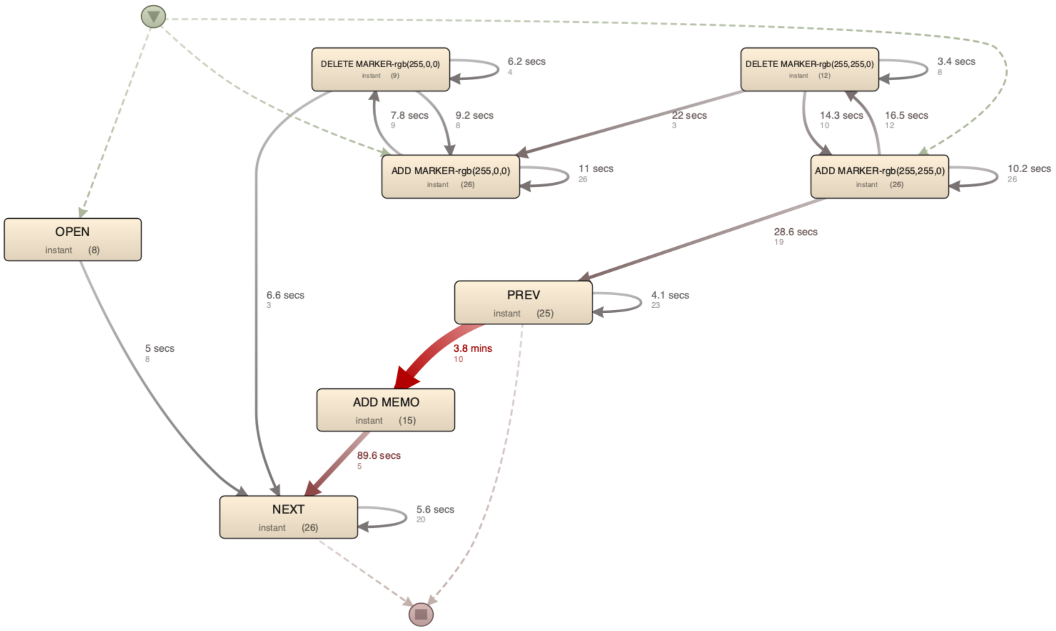


Figure 4. Interaction Pattern of Low Performing Group (n=26).

Low performers mostly highlighted the easily identifiable markers. More importantly, their memos mentioned the obvious points of cultural references and performative elements. Memos lacked synthesis of various references in different pages of the text and how they cumulatively add up overall. One example memo reads like this: (Student_id: 1538, score 5.5/10) "*Cultural: The Cultural reference to Lava Kusha, Rama Laxmana and Krishna Balrama and signifies the strength in the bond between the two friends and shows it as unbreakable. Performative: The first line of the play sets up the stage for the play. Most other performative elements describe the movements, emotions or actions of the actors*."

In the above example, the student has pointed out the reference to the text of *Ramayana;* but falls short of dealing with other significant references to the nuances discussed in section 5.1. What one can note from the memo is only a limited and largely surface-level understanding of the text. There seems to be a lack of deeper engagement with the text while attempting the task. The total time spent on the critical reading task in the BookRoll environment was 32 minutes before writing the memo and the pattern of interactions also highlights no time spent in modifying or revisiting the memos.

4.3.2 High Performers Interaction Patterns and Artefacts Created.

For the high performing group the prominent initiating states were either navigation or adding a red marker (see Fig 6). 50.6% of activities happened in pages 8-10. Revisiting and editing memos was a significant action for the high performers which was not present for the low performers.

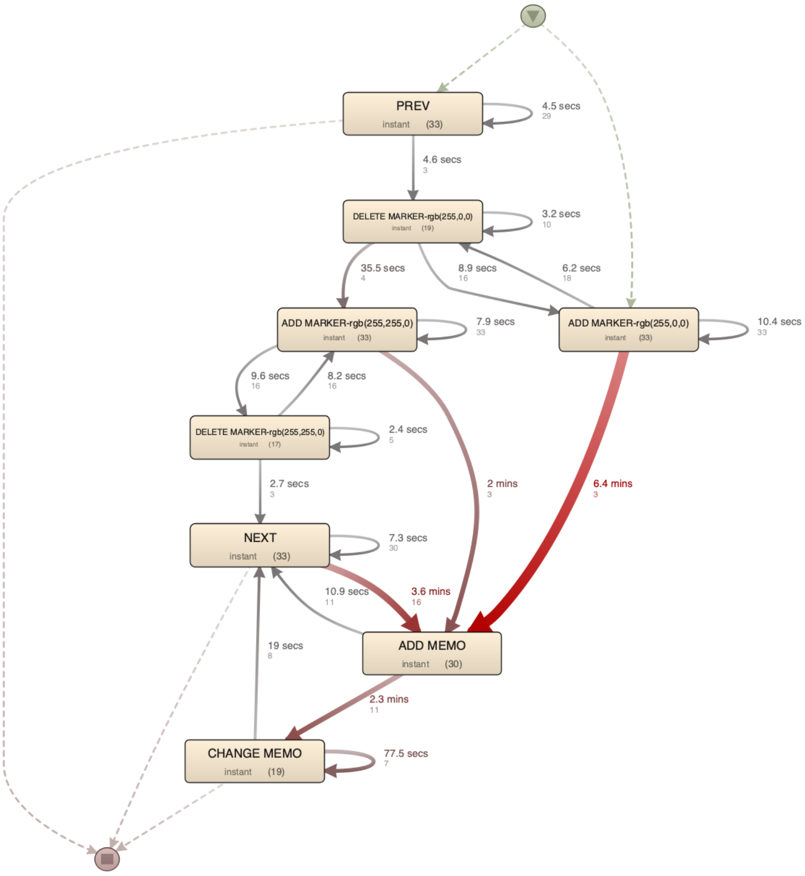


Figure 5. Interaction Pattern of High Performing Group (n=33).

Here is a sample memo by a high performer (student_id: 389, score: 9.5/10) on performative elements: the memo at the very beginning captures the use of masks, music and the significance of these; simultaneously it makes a note of how displaying such vivid imagery keeps the audience's interest. The student is not only capturing the essentials, but is able to move beyond the text and make sense of the intention of the playwright in specific: "*Masks and Music are very instrumental elements of any play, since plays are an inherently different form of literature compared to written works, in the sense that plays involve a lot of dramatization and display vivid imagery to keep the audience entertained and engrossed in them. Here, the music is used by the narrator as an external narrative element to beautifully portray events happening in the story and to provide exposition that keeps the audience always engrossed.*" Further, the student is able to reflect upon the use of songs in this specific context as well as other contexts: "*On closer inspection, we can notice that these songs occur intermittently between the play to rejuvenate and renew the audience's interest as well as provide a means of a cultural form of enjoyment through music's universal appeal to humans.*" Pausing for a while, the memo highlights the narrative function of the songs: "*Also, this particular song sung by the female chorus is actually foreshadowing what will happen to Padmini, Devadatta, and Kapila later in the story.*" Then reflective task is continued as the student adds observations on the point about masks and costumes used in the play: *Along with striking audio imagery provided by music, plays also provide striking visual imagery through the use of masks for actors and eye catching costumes.* Before completing the memo, the student now is able to reflect upon the aesthetic and narrative/ dramatic functions of the use of masks: "*These masks come in myriad designs, each one perfectly describing the nature and behavior of the character wearing it. For example, here we have Devadatta wearing a pale white mask to reflect his soft, gentle nature and build, and we see Kapila wearing a dark black mask which perfectly reflects his strength and might.*"

This memo by a high performer is a carefully observed, comprehended, revisited, and synthesized reading of the text; it also exhibits his/her ability to articulate the ideas cogently. The interaction process in this particular case indicates the student revisited, modified and added memos at various points in the timeline. Reflective reading, conceptual clarity and finesse in the articulation are evinced in the performance of this particular performer who spent 54 minutes in BookRoll and updated the 2 memos 5 times in total before submitting it.

5. Discussions and Conclusions

5.1 Identifying Cultural Reference and Performative Elements during OBE.

Hayavadana, the play is deeply rooted in the Indian cultural milieu, and since a typical student in this study context probably had enough exposure to Indian culture, cultural references are easier to identify. The first few pages have several references and beliefs common to Indian society- irrespective of one's religious affiliation. References to 'Lord Ganesha', 'Vighneshwara, the destroyer of obstacles' , 'husband of Riddhi and Siddhi', 'Goddess Kali' etc. are easy to identify as cultural references for the target audience in this case. Also, there are references to various myths and folk tales from India, the 'Princess of Karnataka', 'Gandharva', 'Goddess of Chitrakoot', reference to places of pilgrimage for the Hindus etc. some of which require a knowledge of the rich mythological texts from India. Regarding performative elements, the play Hayavadana draws from contemporary theatrical conventions and also harks back upon Indian theatrical roots, especially from the Yakshagana form of folk theatre which is new for most of the students. Easier choices to highlight in this context are the frequent stage directions in the play. The highlights on Bookroll indeed confirm it; points such as 'The stage is empty except for a chair...', ' The Bhagavata sings verses in praise of Ganesha', 'The actor goes out' etc. are heavily highlighted while the more nuanced references to the theatrical conventions are identified by very few students. Reference to the audience, actors addressing the audience directly at times, and use of other such metatheatrical conventions such as the use of masks, dolls, half curtains etc. are identified and highlighted by a smaller number of students among the high performers.

5.2 Contributions of the Current Finding and Implications for Technology Design

The reading logs of interactions in BookRoll system and the processed data from the learning analytics dashboard were used to investigate two specific research questions about the relationship of reading behavior to OBE achievement and about the differences in the interaction behaviors and artefact created by the high and low achievement groups. We found that the total duration of the editing action was significantly correlated to the score for this specific time. In previous work (Majumdar 2021), four different reader's profiles emerged during an in-course non-evaluative critical reading activity: Effortful, Strategic, Wanderer and Check-outs. From that perspective most of the reading attempts during the OBE would be either Effortful or Strategic. The current study further analysed interaction sequences during an OBE session about which we did not find any prior literature. Such an analysis was afforded by the LEAF technology framework that logged the students' interactions and limited previously. We differentiated high and low group interaction patterns (compare Fig 5 and 6) and artefacts generated during OBE. High performers had significant reflective states such as changing the memos, cycles of deleting and recreating the markers in their overall interaction process. Such behaviors were limited in the case of low performers. It confirms the initial correlation of editing duration and the achievement in the OBE.

5.2 Limitations and Future Work

The current study investigated the relationship of reading behaviors and achievement in OBE from interaction log data and students' artefacts probably for the first time. However there are certain limitations that require further attention. The overall process of attempting the OBE is captured through log data which does not capture the prior conceptual understanding of the students, their motivation or any of their prior dispositions. These might have implications of how the student is engaging in the task and their behaviors. Further studies are required to investigate differences in student's online reading behavior of the same text prior to exam and during exam.

The BookRoll portal not only enabled tracing the student's reading behavior, but also assisted the instructor to easily upload the reading materials and instructions and the student to answer in the same portal. The instructor could check the students answers provided in the memos directly in the Analysis tool and export the list for the whole class for grading. While this enabled the workflow of conducting the OBE smoother during the COVID19 emergency remote teaching, further technical support can also be developed to automatically evaluate the OBE task given in the current context. Highlighting through markers and writing memo actions of learners can potentially be used to give them feedback. During the data analysis process, the instructor highlighted the portions of the text for reference. Currently we are preparing to present the instructor's highlighted part in the analysis dashboard for the learners to check.

Acknowledgements

This research was supported by JSPS KAKENHI 16H06304, 20K20131, 20H01722, SPIRITS 2020 of Kyoto University.

References

Ashri, D., & Sahoo, B. P. (2021). Open Book Examination and Higher Education During COVID-19: Case of University of Delhi. *Journal of Educational Technology Systems.* https://doi.org/10.1177/0047239521013783

Ben-Yehudah, G., & Eshet-Alkalai, Y. (2018). The contribution of text-highlighting to comprehension: A comparison of print and digital reading. *Journal of Educational Multimedia and Hypermedia,* Vol. 27(2), pp. 153-178.

Block, R. M. (2012). A discussion of the effect of open-book and closed-book exams on student achievement in an introductory statistics course. *Primus*, Vol. 22(3), pp. 228-238. DOI: 10.1080/10511970.2011.565402

Chan C.(2009) Assessment: Open-book Examination, Assessment Resources@HKU, University of Hong Kong [http://ar.cetl.hku.hk]: Accessed on: May 2021

Holm, P., Jarrick, A., & Scott, D. (2015). Humanities world report 2015. Springer Nature.

https://fluxicon.com/disco/ [Computer software]

Improbable achievement; BITS-a profile of change Birla Institute of Technology and Science, Pilani (1990) Birla Institute of Technology and Science, Pilani New Delhi: Wiley Eastern 1990

Jan Larsson & Inger Holmström (2007) Phenomenographic or phenomenological analysis: does it matter? Examples from a study on anaesthesiologists' work, *International Journal of Qualitative Studies on Health and Well-being*, Vol. 2(1), pp. 55-64, DOI:10.1080/17482620601068105

JASP Team (2019). JASP (Version 0.11.1) [Computer software] Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., Ly, A., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Wild, A., Knight, P., Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2019). JASP: Graphical Statistical Software for Common Statistical Designs. *Journal of Statistical Software*, Articles, 88(2), 1–17.

Karnad, G. (1972) Havayadana, in The Oxford Dictionary of Plays. online reference https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095925443

Majumdar R., Bakilapadavu G., Majumder R., Mei-Rong C. A., Flanagan B. and Ogata H. (2020) Learning Analytics of Critical Reading Activity: Reading Hayavadana during Lockdown. *in Procs. of ICCE 2020,* Vol.1, pp. 127-136.

Majumdar R., Bakilapadavu G., Majumder R., Mei-Rong C. A., Flanagan B. and Ogata H. (2021) Learning Analytics of Humanities Course: Reader Profiles in Critical Reading Activity. *Research and Practice in Technology Enhanced Learning*. 16, 25. https://doi.org/10.1186/s41039-021-00164-w.

Ogata H., Majumdar R., Akçapınar G., Hasnine M.N., Flanagan B., Beyond Learning Analytics: Framework for Technology-Enhanced Evidence-Based Education and Learning, *in Procs of the ICCE2018*, pp. 486-489,

Ogata, H., Oi, M., Mohri, K., Okubo, F., Shimada, A., Yamada, M., Wang, J., & Hirokawa, S. (2017). Learning analytics for e-book-based educational big data in higher education. *In Smart Sensors at the IoT Frontier*, Springer, Cham, pp. 327-350.

Rakes, G. C. (2008). Open book testing in online learning environments. *Journal of Interactive Online Learning,* 7(1), pp. 1-9.

Theophilides, C., & Koutselini, M. (2000). Study behavior in the closed-book and the open-book examination: A comparative analysis. *Educational Research and Evaluation*, 6(4), 379-393.

Tor Vidar Eilertsen, Odd Valdermo(2000) Open-book assessment: A contribution to improved learning?, *Studies in Educational Evaluation*, Vol. 26, Issue 2, 2000, pp. 91-103, ISSN 0191-491X.

Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Human Factors in the Adoption of M-Learning by COVID-19 Frontline Learners

Ryan EBARDO[a*] & Merlin Teodosia SUAREZ[b]
[a]*José Rizal University, Philippines*
[b]*De La Salle University – Manila, Philippines*
*ryan.ebardo@jru.edu

Abstract: Mobile Learning is crucial to the continuity of healthcare education during COVID-19. Despite its penchant for the traditional delivery of course content through classroom and clinical settings, M-Learning proved to be a viable solution in a pandemic due to social isolation, community restrictions, and safety concerns. We invited 219 frontline learners from 3 universities, active healthcare professionals who are currently enrolled, to test a structural model based on the Theory of Reason Action. We positioned the human factors of cognitive, social, and affective needs as determinants of attitude in the behavioral intention to adopt M-Learning. We further hypothesize that social norms positively influence the behavioral intention to adopt M-Learning among healthcare frontliners. We applied PLS-SEM to analyze the survey data and revealed that human factors positively influence attitude, leading to the behavioral intention to adopt M-Learning. Social norms and their influence on the behavioral intention to adopt this technology are not supported. We discuss the implications of our study, acknowledge its limitations while mapping out directions for future works to understand M-Learning adoption further.

Keywords: M-Learning, human factors, COVID-19, medical education, healthcare education

1. Introduction

Mobile learning or the use of mobile devices to acquire new knowledge revolutionized 21st-century education. Improvements in mobile technologies and more comprehensive Internet connectivity allowed learners to acquire new knowledge anytime and anywhere through Mobile learning or M-Learning, a benefit that prior research identified as a primary motivation in its increased adoption (Baghcheghi et al., 2020; Maharsi, 2018). The ubiquity of mobile devices, along with a better understanding of the scholarship in its integration with various learning processes, brought forward significant improvements in M-Learning such as context-sensitivity, improved interaction features, and personalization (Lall et al., 2019; Senaratne & Samarasinghe, 2019).

While research on the diffusion of M-Learning in higher education abounds, its adoption in healthcare professional education appears lacking. Prior research has focused primarily on the contexts of university students revealing that today's generation of university learners prefer acquiring new knowledge in the mobile environment (Baghcheghi et al., 2020; Gómez-Ramirez et al., 2019; Qashou, 2021). The quality of healthcare is heavily anchored on what our medical professionals know from their experiences, practices, and formal education; therefore, M-Learning has become an essential vehicle in balancing their profession and the need to acquire new knowledge (Qureshi et al., 2020). Despite the potential of M-Learning to the improvement of the healthcare profession, some opportunities are ripe for further scrutinies such as studies from developing economies (Barteit et al., 2020) and the effects of its adoption beyond technology factors (Attalla et al., 2020; Azizi & Khatony, 2019).

The onset of the novel coronavirus 2019 or COVID-19 disrupted the way we deliver education due to challenges imposed by community lockdowns, social distancing, and campus closures (Pokhrel & Chhetri, 2021). In the Philippines, higher educational institutions or HEIs recalibrated their strategies to ensure continuous learning through a mix of blended and online learning delivery modes (Joaquin et al., 2020; Pelmin, 2020). At the forefront of the battle against COVID-19 are the healthcare professionals who risk their lives to ensure that humanity's battle against the current pandemic is sustained. In fulfilling their roles as frontliners in this battle, learning continues, and the challenge to balance their psychological, cognitive, and social needs can be addressed by M-Learning (Cedeño et al., 2021).

In this study, we approached frontliners – medical doctors and allied professionals – who are currently enrolled in a graduate healthcare management program in three (3) universities in the Philippines. We developed a quantitative model based on the Theory of Reasoned Action with the addition of human factors of cognitive, affective, and social aspects to investigate their influence in the behavioral intention to adopt M-Learning. Given that studies in technology-enabled learning are context-driven, we contribute to current literature in three areas: adoption of M-Learning among adult healthcare students in a developing economy, effects of non-technology factors in M-Learning adoption, and understanding M-Learning during COVID-19 adoption from the perspective of frontline healthcare professionals (Barteit et al., 2020; Freedman & Nicolle, 2020; Heinze & Hu, 2009; Negrescu & Caradaica, 2021). In the following sections, we provide an overview of recent literature in the use of M-Learning in healthcare education, discuss the theoretical underpinnings of our study, present our methodology, highlight the results of our study and conclude by mapping out our recommendations for future research.

2. Related Studies and Theoretical Foundations

Healthcare education has long been viewed as a discipline heavily reliant on the traditional delivery of education. Before COVID-19, medical education relies heavily on practical knowledge application where lessons are delivered mainly within a classroom or a hospital. This practice is also driven by the preference of attending physicians who practice within the clinical settings as there is a need for students to interact, a challenge to which M-Learning is still grappling to address (Li & Bailey, 2020). However, recent literature has reflected the value of disruptive technologies in education demonstrating the viability of M-Learning in healthcare education (Qureshi et al., 2020).

In the literature of technology adoption, attitude and social norms play an essential role in the behavioral intention in using a specific technology. The Theory of Reasoned Action, or TRA, posits that a positive attitude towards technology will facilitate eventual adoption (Ajzen, 1975). The decision to use this theory is guided by prior information system research that found TRA flexible to incorporate external variables and its applicability to the M-Learning adoption domain (Attalla et al., 2020; Buabeng-Andoh, 2018). Like attitude, social norms are also a strong predictor of technology adoption. The likelihood of adopting technology is highly influenced by social pressure. In M-Learning, these two factors have been proven to predict its adoption effectively. For example, studies among university students found that attitude is a dominant predictor of the behavioral intention to adopt M-Learning (Buabeng-Andoh, 2018; Qashou, 2021). Social pressure exerts a certain level of influence when deciding whether to perform a specific behavior. In the context of technology adoption in education, this social pressure may come from classmates and is a strong determinant in the behavioral intention to adopt technology (Raza et al., 2018). Given that a favorable attitude towards M-Learning and those social norms are strong predictors of the behavioral intention in its adoption, within the context of TRA, we propose the following hypotheses:

H1: ***Attitude has a positive and significant influence on the behavioral intention to adopt M-Learning***

H2: ***Social norms have a positive and significant influence on the behavioral intention to adopt M-Learning***

COVID-19 fostered renewed attention to M-Learning as it has become a tool to sustain learning, especially in healthcare education (Cruz-Cunha & Mateus-Coelho, 2021; Rose, 2020). In prior literature, the experiences of medical students were found to be positive towards technology-supported learning during the pandemic (Alsoufi et al., 2020). During COVID-19, M-Learning proved to be a viable solution to learning disruptions due to its various strengths, such as flexibility, asynchronous features, automated class management, and speed (Cedeño et al., 2021; Juan et al., 2020). While research viewed healthcare learners as a cohort who learns best within the clinical settings, the sudden shift to the M-Learning environment may pose challenges in its adoption and requires further inquiry.

COVID-19 may present a unique context where online learners experience heightened psychological stress and increased social isolation, impacting the way they learn online (Brand, 2020). While technology factors in innovation adoption inspired prior literature, human factors and their effects on technology adoption are equally important (Attalla et al., 2020; Roberts & Flin, 2019). In adult learning, several factors drive M-Learning adoption. Adult learners have cognitive needs to grow professionally, and during the COVID-19 pandemic, opportunities paved the way for healthcare professionals to learn new skills through M-Learning (Pokhrel & Chhetri, 2021; Wayne et al., 2020). Aside from cognitive needs, adult learners are driven by the need to interact and socialize, whether within their professional networks or their significant others (Huang, 2016). Current platforms of M-Learning improved their features to integrate ways for better interactivity to achieve social presence. While cognitive and social needs can influence the way adult learners adopt M-learning, the psychological impact of COVID-19 can influence the learning process, especially among healthcare professionals (Brand, 2020; Cedeño et al., 2021). Adults learn best through experiential learning, which is prominent in medical education (Jin et al., 2019). Through M-Learning, adult learners participate in a group where they can disclose, share and discuss emotional distress that can potentially address their affective needs (Tang & Hew, 2018). Given that cognitive, social, and affective needs are factors that can positively influence the attitude of frontline learners towards M-Learning during COVID-19, in the context of our study, we propose the following hypotheses:

H3: ***Cognitive needs has a positive and significant influence on the attitude towards M-Learning***
H4: ***Social needs has a positive and significant influence on the attitude towards M-Learning***
H5: ***Affective needs has a positive and significant influence on the attitude towards M-Learning***

In recent literature on M-Learning in healthcare education, studies have shown that TRA constructs are established predictors of the behavioral intention in its adoption. Attitude or the positive feelings about M-Learning (H1) and social norms, or the perceived influence of significant others (H2), can influence healthcare students' adoption of M-Learning. COVID-19 presents a unique context, and in the adoption of, M-Learning we hypothesized that cognitive (H3), social (H4), and affective needs (H5) are salient considerations in technology. While a few studies have integrated these factors in the behavioral intention to adopt M-Learning, these studies are mostly limited before COVID 19 and using a cohort of medical students at the university level. We summarize our five (5) hypotheses in Figure 1 – Structural Model.

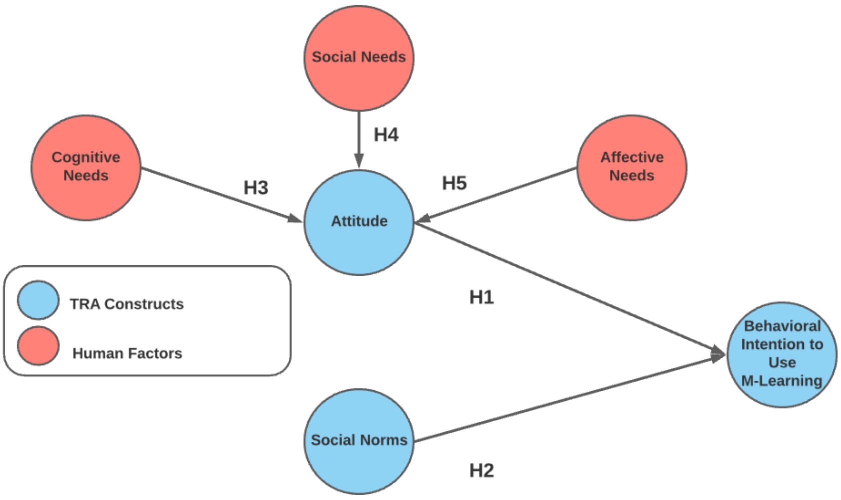


Figure 1. Structural Model.

3. Methodology

We approached three (3) HEIs currently offering a hospital management graduate degree in partnership with a healthcare professional management society to test the structural model. Students enrolled in the program are employed healthcare professionals who are physicians, healthcare administrators, nurses, laboratory staff, and other allied professionals. A total of two hundred nineteen (219) respondents provided their informed consent and answered an online survey through Google Forms.

3.1 Instrument Development

To operationalize the constructs of our structural model, we combined questions from the instrument of Hashim et al. (2014) in their study on adult learners' adoption of M-Learning and the instrument of Huang (2016) in a study investigating social factors in the continuous intention to use technology-based learning. We added questions related to demographics such as gender, age, area of practice, and devices used for M-Learning. An explanation of the study's objectives, the definition of M-Learning, and sample activities were stated at the beginning of the survey for further contextualization. The final version of the instrument consists of twenty-six (26) questions. We invited four students (4) to answer the survey to get their initial feedback. Minor modifications were made, such as the addition of Instant Messaging apps as an example of a communication tool in the social need construct and reframing questions in the social norms construct to mobile devices for further contextualization.

3.2 Validity and Reliability Tests

To further validate the instrument, we purposively selected thirty-one (31) students from the three participating HEIs to answer the survey as a pilot test. A Partial Least Squares or PLS algorithm was applied to the initial results using SmartPLS. Specifically, this test will ensure that the questions or indicators accurately represent the constructs in our structural model. The validity and reliability tests using the PLS algorithm are shown in Table 1 – Instrument Validation. The lowest scores for the Cronbach's Alpha and Composite Reliability or CR measures are 0.789 and 0.862. Given that these scores meet the minimum threshold of 0.70, the instrument demonstrates satisfactory internal consistency. On the other hand, the lowest score for the AVEs is 0.610, which meets the minimum threshold of 0.50, thereby exhibiting adequate convergent validity.

Table 1. *Instrument Validation*

Construct	Cronbach's Alpha	Composite Reliability	Average Variance Extracted
Cognitive Needs	0.830	0.872	0.657
Social Needs	0.789	0.862	0.610
Affective Needs	0.811	0.886	0.631
Attitude	0.870	0.921	0.795
Social Norms	0.928	0.954	0.874
Intention	0.941	0.962	0.894

3.3 Discriminant Validity

The discriminant validity scores check the presence of a high correlation among the constructs of a structural model. It ensures that a specific construct has a unique explanatory power. We extracted the Fornell-Larcker criterion test score from the PLS algorithm in the prior section to test discriminant validity, as shown in Table 2 – Fornell-Larcker Discriminant Validity Test. The diagonal values highlighted in bold text indicate the highest scores compared to non-diagonal values for each construct, demonstrating the absence of inter-correlation and establishing strong discriminant validity for each variable.

Table 2. *Fornell-Larcker Discriminant Validity Test*

Construct	Affective Needs	Attitude	Cognitive Needs	Intention	Social Needs	Social Norms
Affective Needs	0.794					
Attitude	0.574	0.892				
Cognitive Needs	0.556	0.299	0.810			
Intention	0.547	0.792	0.417	0.945		
Social Needs	0.665	0.662	0.559	0.616	0.781	
Social Norms	0.626	0.694	0.367	0.627	0.547	0.935

While the Fornell-Larcker discriminant validity test has been used in prior information systems research and found to be sufficient, recent literature highlighted its reliance on factor loading estimates necessitating a further test using Heterotrait-Monotrait or HTMT test (Hamid et al., 2017; Hair et al., 2017). We extracted the HTMT criterion scores from the PLS algorithm as shown in Table 3 - Heterotrait-Monotrait Validity Test. All values are below 0.85 except for the HTMT score of attitude and intention, which is 0.874. Traditionally, HTMT scores of 0.85 indicate discriminant validity. However, recent updates to the PLS method as applied in IS research have deemed values below 0.90 acceptable (Benitez et al., 2020). The results of the Fornell-Larcker and HTMT tests demonstrate strong evidence that the constructs can represent the dimensions of our structural model and are sufficient to accept or reject our given hypotheses.

Table 3. *Heterotrait-Monotrait Validity Test*

Construct	Affective Needs	Attitude	Cognitive Needs	Intention	Social Needs	Social Norms
Affective Needs						
Attitude	0.640					
Cognitive Needs	0.661	0.332				
Intention	0.577	0.874	0.490			
Social Needs	0.801	0.778	0.695	0.700		
Social Norms	0.721	0.771	0.426	0.665	0.623	

3.4 Participants and Test for Common Method Bias

We deployed our online survey from March to May of 2021. All respondents are currently affiliated with a healthcare institution and enrolled in a postgraduate degree in healthcare management. Of the 219 respondents, 123 or 56% are female, and 96 or 44% are male. In terms of age groups, 15 or 7% are between 20 and 29 years old while 72 or 33% are between 30 and 39 years old. Additionally, 38 or 17% belong to the age group of 40-49 years old, while 68 falls into the 50-59 age group. Of the sample, 26 or 12% are considered older adults. Most of the participants, 146 or 67%, are employed within Metro Manila, while 73 or 33% practice their profession in the provinces.

Common method bias or CMB is an ongoing concern, especially in self-reported scales deployed online. It measures the bias in the way respondents answer a survey, the social desirability to finish a survey, or how the words are chosen to gather similar results. To test whether CMB is present in our study, we extracted the inner Variance Inflation Factors as shown in Table 4 – Test for Common Method Bias. There are no VIF values that are greater than 3.3, indicating the absence of CMB.

Table 4. *Test for Common Method Bias*

Construct	Affective Needs	Attitude	Cognitive Needs	Intention	Social Needs	Social Norms
Affective Needs		1.968				
Attitude				1.928		
Cognitive Needs		1.595				
Intention						
Social Needs		1.978				

Social Norms	1.928

4. Discussion of Results

After collecting the survey responses, a Bootstrapping technique using SmartPLS, a structural analysis technique best suited for studies with small sample sizes, was applied (Benitez et al., 2020; Schmidheiny, 2014). Specifically, we were interested in the T-Statistics values for each path to accept or reject a specific hypothesis. The results of this test are presented in Table 5 – Structural Model Test. A T-Statistics value of above 1.96 means that the relationship is significant (Hair et al., 2014).

Table 5. *Structural Model Test*

HYPOTHESIS	SD	T STATISTICS	P Values	DECISION
H1: Attitude has a positive and significant influence on the behavioral intention to adopt M-Learning	0.126	2.372	0.018	Accept
H2: Social norms have a positive and significant influence on the behavioral intention to adopt M-Learning	0.123	1.228	0.220	Reject
H3: Cognitive needs has a positive and significant influence on the attitude towards M-Learning	0.085	2.171	0.030	Accept
H4: Social needs has a positive and significant influence on the attitude towards M-Learning	0.120	4.697	0.000	Accept
H5: Affective needs has a positive and significant influence on the attitude towards M-Learning	0.126	2.372	0.018	Accept

The human factors of cognitive needs (H1), social needs (H2), and affective needs (H3) have a direct and positive influence on the attitude of frontline learners towards M-Learning based on the T-Statistics values of 2.171, 4.697, and 2.372, respectively. These values are above the minimum threshold of 1.96, demonstrating significant relationships between these human factors and attitude, resulting in the acceptance of H1, H2, and H3 (Hair et al., 2014). Like the findings of prior studies in adopting M-Learning, the factors of cognitive needs, social needs, and affective needs affect how learners view this learning modality (Hashim et al., 2014; Lin & Su, 2020). Although investigations in the adoption and usage behaviors of learners in the medical field established a strong preference for knowledge delivery via classroom or clinical settings to meet their cognitive needs (Lall et al., 2019), the restrictions and safety concerns imposed by COVID-19 highlighted the benefits and affordances of M-Learning in healthcare education (Alsoufi et al., 2020; Cedeño et al., 2021; Rose, 2020). Given that the COVID-19 situation is unprecedented, its impact on patient care, hospital operations, and clinical procedures will need to adjust, and information is best delivered through the M-Learning modality due to its speed, flexibility, and convenience. A massive shift towards M-Learning has been observed where urgent findings of COVID-19, best practices, and government policies are delivered via webinars to medical frontliners (Al-Ahmari et al., 2021; Nepal, 2020).

Like the influence of cognitive needs on attitude, social and affective needs shape the perceptions of frontline learners towards M-Learning. The psychosocial needs to socialize and acquire affection are heightened among learners during COVID-19 mainly due to social isolation, stress, and fear (Joaquin et al., 2020; Pokhrel & Chhetri, 2021). Evidence from prior pandemics has stressed that healthcare workers are most vulnerable to the adverse psychological effects of a health crisis and will disrupt the continuity of learning (Brand, 2020). Among frontline learners, opportunities to discuss and socialize with peers and fellow healthcare professionals on the various topics related to COVID-19 can meet their psychological needs and cushion the negative impact of this pandemic (Brand, 2020; Wilcha, 2020). Additionally, synchronous classes delivered via M-Learning allow breakout rooms where students can freely interact with classmates facilitating lost physical, social connections and acquire peer to peer support (Chandler, 2016; Sneddon et al., 2021).

Consistent with prior findings, a positive attitude towards M-Learning leads to the behavior intention of its adoption (Azizi & Khatony, 2019; Raza et al., 2018). The T-Statistics value of 2.372 (H1) infers that it has a direct and positive influence on the intention to use M-Learning among frontline learners (Hair et al., 2014). As discussed in the prior section, aside from meeting the cognitive needs of learners, M-Learning can facilitate social interactions and provide social support, valuable human needs that are important during this pandemic. In the context of this study, the Philippines experienced one of the most prolonged closures of academic institutions, and M-Learning supported the continuity of medical education (Cedeño et al., 2021; Pelmin, 2020). On the other hand, the T-Statistics value of 1.228 for the relationship of social norms and intention to adopt M-Learning (H2) is not supported as it does not meet the minimum value to establish significance (Hair et al., 2014). While it contradicts the other studies (Gómez-Ramirez et al., 2019; Kucuk et al., 2020), it aligns with the study of Azizi and Khatony (2019). Among adult learners, social norms may not necessarily come from classmates but may come from other social networks such as professional communities of practice, family members, and superiors (Hadadgar et al., 2016; Park et al., 2021). In addition, while we find the influence to be positive but not significant, social norms may not necessarily influence students to use M-Learning as it is the only modality that the participating universities currently offer during the COVID-19 crisis. Lastly, a possible explanation is the recent exploration of subjective norms, which argue that others weakly influence adult learners if they have a solid positive attitude towards M-Learning and a high level of cognitive needs (Hossain et al., 2020).

5. Conclusion

In summary, the results of our SEM analysis confirm that the human factors of cognitive needs, social needs, and affective needs are positively related to the attitude, which in turn leads to the behavioral intention to adopt M-Learning. In the context of this study, we also found that social norms have no direct influence on attitude, and influence may come from other sources, given that the participants of the study are adult learners. The study further established the applicability of TRA in IS research and confirms its flexibility to integrate external variables. M-Learning is well-researched, but COVID -19 and the involuntary shift to online modalities renewed calls to synthesize further how education can be best delivered. Given that prior studies emphasized culture and context in IS research, we contribute to the unfolding scholarship on M-Learning adoption during a pandemic through investigating the influence of human factors in its adoption, using participants from the healthcare sector, and presenting a perspective from a developing economy.

The COVID-19 situation presents a dichotomy of a threat and an opportunity for healthcare education (Brand, 2020). The unknown in medical education is always balanced by scientific curiosity. While the ongoing pandemic disrupted how healthcare professionals learn in physical classrooms and clinical settings, it is an opportune time to embrace innovative technologies as a complementary tool in healthcare education. Integrating advances in educational technologies such as virtual reality to address the lack of clinical practice and the use of telemedicine platforms to interact with patients may be a way to move forward (Remtulla, 2020). Another revelation in this study is the importance of humanizing M-Learning. Although there is an urgent need to continue medical education, universities should balance this with empathy where the well-being of learners is also considered and the pedagogical and curricular implications of M-Learning. Possible ways to further humanize M-Learning are using breakout rooms to encourage free and intimate discussions, utilizing interactive discussion boards, integrating social technologies, and implementing self-care academic breaks (Chandler, 2016; Qureshi et al., 2020; Rapanta et al., 2020).

Our study is limited by several research constraints but can guide future scholarly undertakings to understand M-Learning further. Foremost among these limitations is the small sample size. Future research can replicate our study to a randomized, nationally representative sample size to aid policymakers in deploying M-Learning. Second, we focused our attention on healthcare professionals in the Philippines; comparative studies with other countries can further contextualize our study and understand the role of culture in adopting M-Learning in medical education. Third, the quantitative results can be further explained by qualitative inquiries such as interviews or focused group discussions to shed light on the constructs of human factors and how universities can integrate our findings in deploying M-Learning platforms and crafting their academic policies. Lastly, we conducted our study during the COVID-19 pandemic; another study can be conducted after this crisis to verify whether the results are still applicable once universities revert to normalcy. The findings and the future directions of this study can help various stakeholders of healthcare education navigate through the intricacies of M-Learning as we slowly go back to where we are before this pandemic, mindful of the lessons learned in an unprecedented situation such as COVID-19.

References

Ab Hamid, M. R., Sami, W., & Mohmad Sidek, M. H. (2017). Discriminant Validity Assessment: Use of Fornell & Larcker criterion versus HTMT Criterion. Journal of Physics: Conference Series, 890(1), 3–7. https://doi.org/10.1088/1742-6596/890/1/012163

Ajzen, I. (1975). Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research. 6(2), 244. https://doi.org/10.2307/2065853

Al-Ahmari, A. N., Ajlan, A. M., Bajunaid, K., Alotaibi, N. M., Al-Habib, H., Sabbagh, A. J., Al-Habib, A. F., & Baeesa, S. S. (2021). Perception of Neurosurgery Residents and Attendings on Online Webinars During COVID-19 Pandemic and Implications on Future Education. World Neurosurgery, 146, e811–e816. https://doi.org/10.1016/j.wneu.2020.11.015

Alsoufi, A., Alsuyihili, A., Msherghi, A., Elhadi, A., Atiyah, H., Ashini, A., Ashwieb, A., Ghula, M., Ben Hasan, H., Abudabuos, S., Alameen, H., Abokhdhir, T., Anaiba, M., Nagib, T., Shuwayyah, A., Benothman, R., Arrefae, G., Alkhwayildi, A., Alhadi, A., … Elhadi, M. (2020). Impact of the COVID-19 pandemic on medical education: Medical students' knowledge, attitudes, and practices regarding electronic learning. PLoS ONE, 15(11 November), 1–20. https://doi.org/10.1371/journal.pone.0242905

Attalla, S. M., Hanafy, N. A., Akter, M., & Ruhi, S. (2020). Screening of medical students' intention to practice mobile-learning in malaysia. Malaysian Journal of Medicine and Health Sciences, 16(10), 40–45.

Azizi, S. M., & Khatony, A. (2019). Investigating factors affecting on medical sciences students' intention to adopt mobile learning. BMC Medical Education, 19(1), 1–10. https://doi.org/10.1186/s12909-019-1831-4

Baghcheghi, N., Koohestani, H. R., Karimy, M., & Alizadeh, S. (2020). Factors affecting mobile learning adoption in healthcare professional students based on technology acceptance model. Acta Facultatis Medicae Naissensis, 37(2), 191–200. https://doi.org/10.5937/afmnai2002191b

Barteit, S., Guzek, D., Jahn, A., Bärnighausen, T., Jorge, M. M., & Neuhann, F. (2020). Evaluation of e-learning for medical education in low- and middle-income countries: A systematic review. Computers and Education, 145(October 2019). https://doi.org/10.1016/j.compedu.2019.103726

Benitez, J., Henseler, J., Castillo, A., & Schuberth, F. (2020). How to perform and report an impactful analysis using partial least squares: Guidelines for confirmatory and explanatory IS research. Information and Management, 57(2), 103168. https://doi.org/10.1016/j.im.2019.05.003

Brand, P. L. P. (2020). COVID-19: a unique learning opportunity if the well-being of learners and frontline workers is adequately supported. Perspectives on Medical Education, 9(3), 129–131. https://doi.org/10.1007/s40037-020-00596-y

Buabeng-Andoh, C. (2018). Predicting students' intention to adopt mobile learning. Journal of Research in Innovative Teaching & Learning, 11(2), 178–191. https://doi.org/10.1108/jrit-03-2017-0004

Cedeño, T. D., Rocha, I. C., Ramos, K., & Uy, N. M. (2021). Learning Strategies and Innovations among Medical Students in the Philippines during the COVID-19 Pandemic. International Journal of Medical Students, 77–79. https://doi.org/10.5195/ijms.2021.908

Chandler, K. (2016). Using Breakout Rooms in Synchronous Online Tutorials. Journal of Perspectives in Applied Academic Practice, 4(3). https://doi.org/10.14297/jpaap.v4i3.216

Cruz-Cunha, M. M., & Mateus-Coelho, N. (2021). m-Learning and m-Health education applications in SARS-Cov2 Pandemic. Proceedings of 2100 Projects Association Joint Conferences – Vol.X (2020), March.

Freedman, A., & Nicolle, J. (2020). Social isolation and loneliness: the new geriatric giants. Canadian Family Physician, 66, 176–182.

Gómez-Ramirez, I., Valencia-Arias, A., & Duque, L. (2019). Approach to M-learning acceptance among university students: An integrated model of TPB and TAM. International Review of Research in Open and Distance Learning, 20(3), 141–164. https://doi.org/10.19173/irrodl.v20i4.4061

Hadadgar, A., Changiz, T., Masiello, I., Dehghani, Z., Mirshahzadeh, N., & Zary, N. (2016). Applicability of the theory of planned behavior in explaining the general practitioners eLearning use in continuing medical education. BMC Medical Education, 16(1), 1–8. https://doi.org/10.1186/s12909-016-0738-6

Hair, J. F., Hult, G., Ringle, C., & Sarstedt, M. (2014). Partial least squares structural equation modeling (PLS-SEM). In Sage Publisher. https://doi.org/10.1108/EBR-10-2013-0128

Hair, J., Hollingsworth, C. L., Randolph, A. B., & Chong, A. Y. L. (2017). An updated and expanded assessment of PLS-SEM in information systems research. Industrial Management and Data Systems, 117(3), 442–458. https://doi.org/10.1108/IMDS-04-2016-0130

Hashim, K. F., Tan, F. B., & Rashid, A. (2014). Adult learners' intention to adopt mobile learning: A motivational perspective. British Journal of Educational Technology, 46(2), 381–390. https://doi.org/10.1111/bjet.12148

Heinze, N., & Hu, Q. (2009). Why college undergraduates choose IT: A multi-theoretical perspective. European Journal of Information Systems, 18(5), 462–475. https://doi.org/10.1057/ejis.2009.30

Hossain, M. N., Talukder, M. S., Khayer, A., & Bao, Y. (2020). Investigating the factors driving adult learners' continuous intention to use M-learning application: a fuzzy-set analysis. Journal of Research in Innovative Teaching & Learning, ahead-of-p(ahead-of-print). https://doi.org/10.1108/jrit-09-2019-0071

Huang, Y. M. (2016). The factors that predispose students to continuously use cloud services: Social and technological perspectives. Computers and Education, 97, 86–96. https://doi.org/10.1016/j.compedu.2016.02.016

Jin, B., Kim, J., & Baumgartner, L. M. (2019). Informal Learning of Older Adults in Using Mobile Devices: A Review of the Literature. Adult Education Quarterly, 69(2), 120–141. https://doi.org/10.1177/0741713619834726

Joaquin, J. J. B., Biana, H. T., & Dacela, M. A. (2020). The Philippine Higher Education Sector in the Time of COVID-19. Frontiers in Education, 5(October), 1–6. https://doi.org/10.3389/feduc.2020.576371

Juan, A., Frontera, G., Ap, C., Ros, I., Narváez, J., Marí, B., & Jm, N. (2020). MEDITERRANEAN JOURNAL elderly. 31(1), 42–49.

Kucuk, S., Baydas Onlu, O., & Kapakin, S. (2020). A Model for Medical Students' Behavioral Intention to Use Mobile Learning. Journal of Medical Education and Curricular Development, 7, 238212052097322. https://doi.org/10.1177/2382120520973222

Lall, P., Rees, R., Yi Law, G. C., Dunleavy, G., Cotič, Ž., & Car, J. (2019). Influences on the implementation of mobile learning for medical and nursing education: Qualitative systematic review by the digital health education collaboration. Journal of Medical Internet Research, 21(2). https://doi.org/10.2196/12895

Li, H. O. Y., & Bailey, A. M. J. (2020). Medical Education Amid the COVID-19 Pandemic: New Perspectives for the Future. Academic Medicine, E11–E12. https://doi.org/10.1097/ACM.0000000000003594

Lin, X., & Su, S. (2020). Chinese College Students' Attitude and Intention of Adopting Mobile Learning. International Journal of Education and Development Using Information and Communication Technology, 16(2), 6–21.

Maharsi, I. (2018). Developing EFL Students' Learning Reflection and Self-Regulated Learning through Google Classroom. ACM International Conference Proceeding Series, 62–66. https://doi.org/10.1145/3234825.3234841

Negrescu, V., & Caradaica, M. (2021). M-Learning and Security Issues in the Coronavirus Era. Ceeol.Com, March, 6–25. https://www.ceeol.com/search/article-detail?id=928130

Nepal, P. (2020). Eastern Green Neurosurgery. 02(01), 52–55.

Park, S., Kim, B., & Kim, K. A. (2021). Preventive behavioral insights for emerging adults: A survey during the covid-19 pandemic. International Journal of Environmental Research and Public Health, 18(5), 1–10. https://doi.org/10.3390/ijerph18052569

Pelmin, M. (2020). Readings on Coronavirus Disease (COVID-19) and the Higher Education Institution (HEIs) Emergency Preparedness in the Philippines. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3573896

Pokhrel, S., & Chhetri, R. (2021). A Literature Review on Impact of COVID-19 Pandemic on Teaching and Learning. Higher Education for the Future, 8(1), 133–141. https://doi.org/10.1177/2347631120983481

Qashou, A. (2021). Influencing factors in M-learning adoption in higher education. In Education and Information Technologies (Vol. 26, Issue 2). Education and Information Technologies. https://doi.org/10.1007/s10639-020-10323-z

Qureshi, M. I., Khan, N., Ahmad Hassan Gillani, S. M., & Raza, H. (2020). A systematic review of past decade of mobile learning: What we learned and where to go. International Journal of Interactive Mobile Technologies, 14(6), 67–81. https://doi.org/10.3991/IJIM.V14I06.13479

Rapanta, C., Botturi, L., Goodyear, P., Guàrdia, L., & Koole, M. (2020). Online University Teaching During and After the Covid-19 Crisis: Refocusing Teacher Presence and Learning Activity. Postdigital Science and Education, 2(3), 923–945. https://doi.org/10.1007/s42438-020-00155-y

Raza, S. A., Umer, A., Qazi, W., & Makhdoom, M. (2018). The Effects of Attitudinal, Normative, and Control Beliefs on M-Learning Adoption Among the Students of Higher Education in Pakistan. Journal of Educational Computing Research, 56(4), 563–588. https://doi.org/10.1177/0735633117715941

Remtulla, R. (2020). The present and future applications of technology in adapting medical education amidst the COVID-19 pandemic. JMIR Medical Education, 6(2). https://doi.org/10.2196/20190

Roberts, R., & Flin, R. (2019). The psychological factors that influence successful technology adoption in the oil and gas industry. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63(1), 1183–1187. https://doi.org/10.1177/1071181319631105

Rose, S. (2020). Medical Student Education in the Time of COVID-19. JAMA - Journal of the American Medical Association, 323(21), 2131–2132. https://doi.org/10.1001/jama.2020.5227

Schmidheiny, K. (2014). Clustering in the Linear Model. Short Guideds to Microeconometrics - Universitaet Basel, 1–11. https://doi.org/10.1016/j.genhosppsych.2011.03.015

Senaratne, S. I., & Samarasinghe, S. M. (2019). Factors Affecting the Intention to Adopt M-Learning. International Business Research, 12(2), 150. https://doi.org/10.5539/ibr.v12n2p150

Sneddon, S., Stapleton, G., & Huser, C. (2021). Twelve tips for online synchronous small group learning in medical education. MedEdPublish, 10(1), 1–10. https://doi.org/10.15694/mep.2021.000076.1

Tang, Y., & Hew, K. F. (2018). Examining the utility and usability of mobile instant messaging in a graduate-level course: A usefulness theoretical perspective. Australasian Journal of Educational Technology, 35(4), 128–143. https://doi.org/10.14742/ajet.4571

Wayne, D. B., Green, M., & Neilson, E. G. (2020). Medical education in the time of COVID-19. Science Advances, 6(31), 4–6. https://doi.org/10.1126/sciadv.abc7110

Wilcha, R. J. (2020). Effectiveness of virtual medical teaching during the COVID-19 crisis: Systematic review. JMIR Medical Education, 6(2), 1–16. https://doi.org/10.2196/20963

Design Guidelines for Scaffolding Self-Regulation in Personalized Adaptive Learning (PAL) Systems: A Systematic Review

Vishwas BADHE*, Gargi BANERJEE & Chandan DASGUPTA
Indian Institute of Technology, Bombay, India
*vishwasbadhe@gmail.com

Abstract: In the present pandemic time, almost all teaching-learning processes have been shifted to online mode. In this online setting, Personalized Adaptive Learning (PAL) educational technology products have become popular amongst educators. PAL systems are facilitating Ubiquitous learning in learner-centric ways by enabling learners to learn anywhere, anytime, at their own pace, and even in settings where teachers are not present (e.g., at home). Such systems require learners to apply self-regulation skills to achieve learning goals. However, guidelines for designing scaffolds for self-regulation in PAL systems for informal (out of school) settings are not readily available. In this paper, we present a systematic literature review of relevant papers published between 2001 and 2021 to understand what guidelines exist. We then propose a set of guidelines that may form the basis for designing effective scaffolds in PAL systems for informal environments for promoting self-regulated learning. The set of proposed guidelines are mapped with 'cyclical phases model' of self-regulation by Zimmerman which will be helpful for PAL system designers.

Keywords: Self-Regulation, scaffolding, ubiquitous learning environments (ULE), Personalized Adaptive Learning (PAL), literature review

1. Introduction

The positive disruption by ubiquitous learning environments in the educational technology domain is experienced by almost all stakeholders in the last decade. A plethora of different Personalized Adaptive Learning (PAL) solutions are being designed by EdTech companies and these are being used by children in ubiquitous learning environments (ULE) which allows seamless mobility in the learning process and is not tied to a specific location. ULE uses a variety of small, portable electronic devices (e.g., smartphones, tablets, smart wearable gadgets, etc.) to create new learning opportunities for learners to learn anywhere, anytime, and at one's own pace (Karoudis & Magoulas, 2016). Such portable environments have seen an increase in use in the present pandemic situation when all teaching-learning processes have shifted to online mode, irrespective of the level of readiness of parents and children. In this online setting, Personalized Adaptive Learning (PAL) educational technology products have become especially popular amongst educators. PAL systems are facilitating Ubiquitous learning (U-learning) in learner-centric ways by enabling learners to learn anywhere, anytime, at their own pace, and even in settings where teachers are not present (e.g., at home). Such systems adapt content and/or assessments based on the learner's performance and interaction with the given resources. To make this process effective, such ubiquitous and adaptive systems require learners to apply self-regulation skills to achieve the learning goals (Leonor & Alejandro, 2019). Self-regulation is defined as the control that learners have over their cognition, behavior, emotions, and motivation through the use of personal strategies to achieve the goals they have established (Panadero & Alonso, 2014). However, to help learners practice self-regulation, sufficient scaffolding is required. But to design scaffolding to SRL in PAL based informal (out of school) ULE systems settings the required guidelines are not available. This paper analyses relevant publications between 2001 and 2021 to present a systematic literature review of existing guidelines. We then propose a set of guidelines that may form the basis for designing effective scaffolds in PAL-based ULE systems for informal environments for promoting self-regulated learning. The following research questions guide our review process- 1) What are the existing design guidelines for scaffolding self-regulation in PAL-based ubiquitous learning environments?, 2) Which of these guidelines have empirical evidence of their effectiveness? and 3) What are the potential pedagogical and technological operationalization of design guidelines for scaffolding self-regulation in PAL-based ULE for K-12 learners ?

2. Methodology

2.1 Search Keywords & Articles

We identified keywords based on the identified domain to search for potential research papers. The search keywords for ubiquitous environments were “ubiquitous learning environments” or “U-Learning” or “ubiquitous technology in learning”. The search keywords for self-regulation are “Self-regulated learning” or “SRL in children” or “SRL in mathematics” or “SRL for academic achievements”. The search keywords for scaffolding are “support in SRL” or “parental support in SRL” or “parental support for children” or “scaffolding in SRL”. The search keywords for exploring models of self-regulation and its comparative analysis are “SRL models” or “Review of SRL models” or “Comparison of self-regulated learning models”. The scope of this review is limited to the past decade i.e., 2001 to 2021.

We searched research databases such as IEEE xplore, Science Direct, Google Scholar to find the relevant research papers. The following inclusion criteria were applied to select the research papers that were found most relevant to our objectives. We considered the papers that included at least meta-level guidelines or provided empirical evidence for the effectiveness of design guidelines for scaffolding in SRL-enabled PALs. The papers which were suggesting some general scaffolding in the classroom or suggesting some parental scaffolding external to PAL were excluded as per exclusion criteria. These inclusion and exclusion criteria were applied to effectively investigate the design guidelines for scaffolding in the self-regulation process. In the initial search using the specified keywords, we have found a total of 679 research papers. Out of 679, we have selected 24 broadly relevant research papers after applying inclusion criteria. After reviewing those 24 research papers, we have found 7 research papers closely aligned with our review objectives.

3. Literature review

3.1 Mapping with SRL Model

To map the design guidelines to the standard SRL models, this study reviewed different SRL models such as the cyclical phases model by (Zimmerman & Moylan, 2009), (COPES) Winne’s SRL model adapted (Winne & Hadwin, 1998). After getting insights from different SRL models, the cyclical phases model by Zimmerman & Moylan (2009) was selected because it has strong connections with social cognitive theory as compared to the COPES model. Zimmerman’s model (Figure 1) also lays more emphasis on the influence of motivation on self-regulation (Panadero & Alonso, 2014) which is an important parameter for children. On the other hand COPES model has a loosely sequenced learning cycle and it does not depict the social cognitive theory precisely (Rovers et al., 2019). Social cognitive theory is important because SRL occurs through the interdependence of individual aspects, linked to feelings, emotions and thoughts, behavior, and the environment in which students find themselves (Zimmerman, 2000). With the help of social cognitive theory, we can assign meaning to learners’ behavior while the learner is using SRL skills.

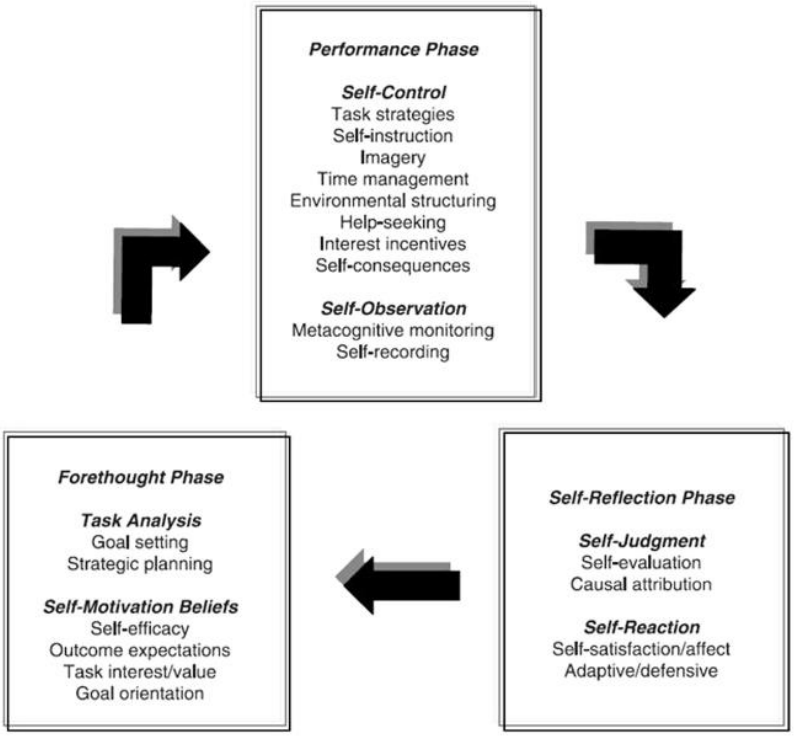


Figure 1. Cyclical Phases Model of Self-regulation according to Zimmerman & Moylan (2009).

Mapping of the existing design guidelines to Zimmerman's model is presented in Table 1.

Table 1. *Existing Design Guidelines for Scaffolding and Its Potential Pedagogical & Technological Operationalization*

Alignment with phases of SRL model by Zimmermann	Design guidelines of SRL	Paper proposing the guidelines	How paper is operationalizing guideline in SRL (supporting arguments for operationalization of SRL)	Potential operationalization for scaffolding SRL in PAL	
				Pedagogical	Technological
Forethought: Goal setting, outcome expectations, task interest/ value, goal orientation	Guideline: have the students set simple, but realistic goals for the pre-class sessions. Support: Have students set a learning goal for a next pre-class session based on their performance in a previous pre-class session	(Yoon et al., 2021)	Allowing the students to set learning goals leads them to initiate the recursive self-regulated learning process toward attaining their ultimate goals. It is important to make the students begin with simple, but realistic goals so that they can have the opportunity to calibrate their goals as they progress. The students' self-efficacy, fostered by their success in prior performances, will, in turn, affect their later self-set goals (Yoon et al., 2021, p. 4)	Computer-based learning environments may facilitate the individual goal setting template in which learners will set their next pre-class learning goals.	Dynamic planning template attached with calendar with timely reminders and appreciation notes. It is the completion of a learning goal which will tell them distance from goal

Forethought: Goal setting, outcome expectations, task interest/ value, goal orientation	Guideline: Foster student motivation by highlighting task values. Support: Have the students ponder ways to transfer what they learned from the pre-class sessions to new contexts	(Yoon et al., 2021; Laere et al., 2015)	Students are required to acquire foundational knowledge by studying the given materials at home without the instructors' explanations as to why the materials are useful (i.e. utility value) and what benefits they can have when they successfully complete the task (i.e. attainment value) (Yoon et al., 2021, p. 5)	From the starting point of the lesson, keep learners informed about the rationale and objective behind learning that topic. (task value)	A kind of pre-test can be designed with open ended answers to reflect upon task value questions
Forethought: Goal setting, outcome expectations, task interest/ value, goal orientation	Guideline: Scaffolding by parents	(Zhang & Whitebread, 2017)	Parents provide adequate metacognitive information in an understandable way and at an appropriate pace along with task-oriented-ness (Zhang & Whitebread, 2017, p. 2)	Support to parents should be provided for the following points: 1) To effectively integrate ULE in their child's learning process at home. 2) To make parents aware of their child's learning behavior to enhance task oriented-ness like encouraging a child to do problem-solving, selection and decision-making	The possible ways a learning system can address the above is to make actionable performance analysis report of their child available to parents in regular intervals like weekly reports along with learning behavioral pattern
Forethought: Goal setting, outcome expectations, task interest/ value, goal orientation	Guideline: Scaffolding to parents	(Muhammad & Iqra, 2020)	Parental autonomy support to learners (Muhammad & Iqra, 2020, p. 2)	PAL in ULE can facilitate opportunities for parents to give autonomy to learners using separate guidelines provided on the parental dashboard	For this, the learning system should give autonomy to learners to choose which chapter, which learning units to do, at what grade level and in what sequence
Performance phase: imagery, metacognitive monitoring	Guideline: Organize instruction and activities to facilitate cognitive and metacognitive processes	(Ley & Young, 2001; Laere et al., 2015)	Overt or covert rearrangement of instructional materials to improve learning (Ley & Young, 2001, p. 2)	Organizing strategies like concept mapping, schematizing (arranging contents in a schematic form) can be implemented	Introduce concept map, concept board or Miro board tool to learners for an activity in which organizing/ relating the content is required

Performance phase: imagery, metacognitive monitoring	Guideline: Use instructional goals and feedback to present the learner with monitoring opportunities	(Ley & Young, 2001)	Record events or results to check its alignment with goals & then feedback can be created (Ley & Young, 2001, p. 2)	Periodic constructive feedback and potential pain points can be highlighted by the PAL and presented before learners. Specific highlights for monitoring can help learners and save their time	Tools for monitoring and checking alignment with goals can be introduced. which checks the consistency of progress with goals and generates correct feedback as required
Performance phase: imagery, metacognitive monitoring	Guideline: Guide learners to prepare and structure an effective learning environment	(Ley & Young, 2001)	Select or arrange the physical setting to make learning easier (Ley & Young, 2001, p. 2)	1) Ask to list the distractions around learners. 2) Advise learners how to arrange physical environments and cope with distractions. 3) Providing a list of strategies will assist less self-regulating learners	Provide a checklist template of guidelines to be followed for setting up the learning environment. This can be followed by the learner
Performance phase: imagery, metacognitive monitoring	Guideline: Help the students accurately monitor their engagements in the pre-class sessions. Support: Use visualizations that show learning activity completion after each pre-class session; use visualizations to show both student progress and performance.	(Yoon et al., 2021)	Allow students to monitor their own learning progress through a support to SRL that visualizes their online behaviors using their log data. The study revealed that the opportunity to obtain information about their learning progress had a positive impact on the students' academic performances (Yoon et al., 2021, p. 4)	1) At the end of each week or fortnight, learners should be provided an opportunity to monitor their own progress so that they can understand their own learning process. 2) After the task, self-report questions	1) Through log data, scores and decisions they have made could be shown using a dashboard so that they can monitor. 2) Tech tool for self-report questionnaire
Self-reflection : self-evaluation, self-satisfaction/ affect, adaptive, defensive	Guideline: Provide learners with continuous evaluation information and occasions to self-evaluate	(Ley & Young, 2001)	Evaluate completed work quality; reread tests to prepare for class (Ley & Young, 2001, p. 2)	Evaluation could not be only comparison between the learner's own performance to a standard, but for comparative outcome between performance and the set standard using them for self-judgment.	Technological options for setting and adjusting evaluation standards and goals can help.

Self-reflection : self-evaluation, self-satisfaction/ affect, adaptive, defensive	Guideline: Scaffold design guidelines: 1) Diagnosis, 2) Calibrated support, 3) Fading, 4) Individualization (Personalized & Adaptive scaffolds)	(Azevedo & Hadwin,2005; Chen, 2014)	Identifying needs and providing correct support till the learner gets mastery (Azevedo & Hadwin, 2005, p. 5)	Continuous diagnosis is required to calibrate/adjust the support of learners.	Remedial scaffolding agents: visualizations, animations, videos, games etc.

Research questions 1 and 2 are addressed in the above table (Table 2) by giving existing design guidelines for scaffolding self-regulation in PAL-based ULE. The first 2 columns are addressing RQ 1 by providing existing design guidelines and mapping with Zimmerman's SRL model. To address RQ 3, in the last two columns, we have tried to propose potential pedagogical and technological operationalization of design guidelines for scaffolding self-regulation in PAL-based ULE for K-12 learners. All research papers reviewed have given empirical evidence for the effectiveness of design guidelines that addressed RQ 2.

The design guidelines for scaffolding SRL given in table 2 above are mapped with Zimmerman's cyclical phases model. Those 3 broad phases are forethought, performance & self-reflection. These broad phases are further divided into micro-level sub-processes. In the forethought phase, learners analyze tasks, set goals, and plan to achieve the goal using different motivational strategies. In the performance phase, the learner executes tasks and monitors one's own progress, and further uses self-control strategies. In the self-reflection phase, learners assess one's own performance and attribute it to levels of success (Panadero, 2017).

4. Recommendations for pedagogical and technological scaffolding in PAL

Based on the literature survey highlighted in table 1, the recommendations for pedagogical & technological scaffolding for PAL-based ULE systems are given in the following table 2.

Table 2. *Pedagogical & Technological Scaffold Recommendations for PAL*

Alignment with phases of SRL model by Zimmermann	Existing Design Guidelines for SRL	Pedagogical & technological scaffolding recommendations for PAL
Forethought: Goal setting, outcome expectations, task interest/ value, goal orientation	Guideline: Have the students set simple but realistic goals for the pre-class sessions. Support: have students set a learning goal for the next pre-class session based on their performance in a previous pre-class session.	Enable adjustment of learning goals in ULE: Facilitate goal setting option for learners using dashboard in PAL ULE.
Forethought: Goal setting, outcome expectations, task interest/ value, goal orientation	Guideline: Foster student motivation by highlighting task values. Support: have the students ponder ways to transfer what they learned from the pre-class sessions to new contexts.	Provide an opportunity to transfer learning to real-life: Facilitate knowledge construction using a constructivist approach and facilitate objectively correct formative assessments for learners.
Forethought: Goal setting, outcome expectations, task interest/ value, goal orientation	Guideline: Scaffolding by parents	Provide Learner autonomy: Equip learner with autonomy support to select learning content as per his/her goal and perception towards task value.

Forethought: Goal setting, outcome expectations, task interest/ value, goal orientation	Guideline: Scaffolding to parents	Provide Parent support: Provide support for parents by which effective adoption of PAL in their child's learning becomes easy and parents could know the learning behavior of their child.
Performance phase: imagery, metacognitive monitoring	Guideline: Organize instruction and activities to facilitate cognitive and metacognitive processes	Provide Instructions in Constructive Approach: Facilitate instructions in a constructive way so that learners could build upon previous knowledge
Performance phase: imagery, metacognitive monitoring	Guideline: Use instructional goals and feedback to present the learner with monitoring opportunities	Provide Motivational features: Include features that motivate the learner intrinsically and extrinsically to explore the content
Performance phase: imagery, metacognitive monitoring	Guideline: Guide learners to prepare and structure an effective learning environment	Provide supporting guideline to structure the environment: To structure one's own learning environment and cope with distractions (checklist)
Performance phase: imagery, metacognitive monitoring	Guideline: Help the students accurately monitor their engagements in the pre-class sessions. Support: Use visualizations that show learning activity completion after each pre-class session; use visualizations to show both student progress and performance	Provide planning & monitoring for self-regulated learning : Provide sufficient guidance for planning and monitoring one's own learning on the PAL dashboard
Self-reflection-self-evaluation, self-satisfaction/affect, adaptive defensive	Guideline: Provide learners with continuous evaluation information and occasions to self-evaluate	Provide continuous real-time information for complementing self-evaluation: To reflect on one's learning, the technological dashboard could provide real-time information about learning in self-evaluation
Self-reflection-self-evaluation, self-satisfaction/affect, adaptive defensive	Guideline: Scaffold design guidelines: 1) Diagnosis, 2) calibrated support, 3) fading, 4) individualization	Provide self-reflection scaffolding to the learner through design: To gain the correct understanding of the topic and relate it to real-life situations. But remove scaffolding slowly when learners get mastery in a skill

5. Discussion

In this paper, we have reviewed prior research on scaffolds for facilitating self-regulated learning in PAL-based ULEs. The design guidelines for scaffolding self-regulation in PAL-based ULEs that emerged from this work show that such scaffolds should provide learners autonomy to select learning content as per their goal and perception towards the value of the task. Sufficient guidance for planning and monitoring one's own learning is imperative in such systems. This can be implemented in the form of a dashboard with the real-time progress of the learner. Such real-time dashboards could nudge learners to become more responsible, self-regulated, and autonomous learners. In addition, with the aid of such dashboard scaffolds, learners will likely gain a deeper conceptual understanding of the topic as they can deliberate on which parts to should pay more attention to. Furthermore, instructions delivered in a constructive way can help learners build on their previous knowledge and relate to real-life situations. Along with that, motivational support and increased involvement of parents with the help of integrated parent modules can help learners engage in self-regulated learning.

One of the unique contributions of this work is mapping the guidelines to different phases of SRL based on Zimermann's cyclical phases model. Thus, these guidelines inform the ULE designers of the scaffolds required for supporting the flow of SRL. Some guidelines may appear counterintuitive to PAL-based ULE systems but these are important to consider as well. For example, support for parents is essential for making them aware of their child's learning behavior and to effectively integrate ULE into the child's learning process. Such parental supports will likely lead to better self-regulated learning amongst the learners. These guidelines can act as a stepping stone to decide what pedagogical and technological recommendations to include in a ULE. For instance, dashboards can have actionable information on learner progress and performance. In addition, they can also have additional sections to address planning and monitoring of goals attained.

6. Conclusion

PAL-based ULEs are systems that are meant to model an individual learner's learning behavior and construct a personalized learning path on behalf of the learner. Yet learners need to be provided a level of autonomy to set their goals along with provision for re-adjustment of their goals after self-reflection on their performance. This underlines the importance of the above set of guidelines for ULE system designers. These guidelines provide the foundation on which further detailed scaffolding features and standards can be developed. Scaffolding self-regulation in ubiquitous learning environments (ULE) will help learners learn anywhere, anytime, and also in the most suitable way to achieve their learning goal. This could be a way forward to cope with learning losses due to school closure and lack of teacher-learner interaction happening in the current pandemic situation.

Acknowledgments

We would like to thank all the people who have supported the process of review and paper writing.

References

Karoudis, K. & Magoulas, G. (2016). Ubiquitous Learning Architecture to Enable Learning Path Design across the Cumulative Learning Continuum. *Informatics, 3*(19), 10.3390/informatics3040019.

Leonor, A. C. R. & Alejandro, P. (2019). A holistic self-regulated learning model: A proposal and application in ubiquitous-learning. *Expert Systems with Applications, 123*, 299-314.

Panadero, E. & Alonso-Tapia, J. (2014). How do students self-regulate? Review of Zimmerman's cyclical model of self-regulated learning. *Anales de Psicología, 30*, 450-462.

Panadero, E. (2017). A Review of Self-regulated Learning: Six Models and Four Directions for Research. *Frontiers in Psychology, 8*, 10.3389/fpsyg.2017.00422.

Rovers, S., Clarebout, G., Savelberg, H., de Bruin, A., & Van Merrienboer, J. J. G. (2019). Granularity matters: comparing different ways of measuring self-regulated learning. *Metacognition and Learning, 14*, 10.1007/s11409-019-09188-6.

Van Laere, E., McKenney, S., & van Braak, J. (2015). Design guidelines for computer-based learning environments aimed at fostering knowledge acquisition in linguistically diverse contexts. *Technology Research & Development*.

Chen, C. (2014). An adaptive scaffolding e-learning system for middle school students' physics learning. *Australasian Journal of Educational Technology, 30*(3), 342–355. doi: 10.14742/ ajet.430

Azevedo, R., & Hadwin, A. F. (2005). Scaffolding self-regulated learning And Metacognition – implications for the design of Computer-based scaffolds. *Instructional Science, 33*(5-6), 367-379. doi:10.1007/s11251-005-1272-9

Ley, K. & Young, D. (2001). Instructional principles for self-regulation. *Educational Technology Research and Development, 49,* 93-103. 10.1007/BF02504930.

Muhammad, S. F. & Iqra, A. (2020). Parental Involvement as Predictor for Self-regulated Learning and Academic Achievement of Students at Secondary School Level. *Journal of Educational Sciences & Research*, *7*(1), 14-32.

Yoon, M., Hill, J. & Kim, D.(2021). Designing supports for promoting self-regulated learning in the flipped classroom. *Journal of Computing in Higher Education*. 10.1007/s12528-021-09269-z.

Zhang, H. & Whitebread, D. (2017). Linking parental scaffolding with self-regulated learning in Chinese kindergarten children. *Learning and Instruction, 49*, 121-130. 10.1016/j.learninstruc.2017.01.001.

Zimmerman, B. J. (2000). Attaining self-regulation. A social cognitive perspective. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), Handbook of self-regulation (pp. 13-39). San Diego, Ca: Academic Press.

Zimmerman, B. J., & Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In D. J. Hacker, J. Dunlosky & A. C. Graesser (Eds.), Handbook of Metacognition in Education (pp. 299-315). New York: Routledge.

Winne, P. H., and Hadwin, A. F. (1998). "Studying as self-regulated engagement in learning," in Metacognition in Educational Theory and Practice, eds D. Hacker, J. Dunlosky, and A. Graesser (Hillsdale, NJ: Erlbaum), 277–304.

Comparison of Experts and Novices in Determining the Gravitational Acceleration using Mobile Phone with Phyphox Application

Aungtinee KITTIRAVECHOTE* **& Thanida SUJARITTHAM**
Program of General Science, Faculty of Education, Bansomdejchaopraya Rajabhat University, Bangkok 10600, Thailand
*aungtinee.ki@bsru.ac.th

Abstract: With the advancement of technology, mobile phones have become the preferred instrument for monitoring a variety of phenomena in the science classroom. There are numerous applications designed to measure the magnitude of gravitational acceleration, Phyphox being one that has yet to be tested. In this study, we comprehensively assess the reliability of Phyphox application through 35 pre-service students (novices) and 5 experts by investigating the experimental mean gravity and limitations and response, specifically the differences between novices and experts in determining gravitational acceleration using a mobile phone with Phyphox and three mathematical analyses: taking average after formula substitution, manual plot, and commercial plot. The results showed that the experimental values of gravitational acceleration done by novices and experts match the theoretical value reported by the National Institute of Metrology (Thailand), with no significant difference in performance between them at .10 level. However, when comparing novices' performance across three mathematical analyses, novices' responses for manually plot differ significantly from those obtained for taking average after formula substitution and commercial plot at .05 level, which is due to some constraints related to novice plotting skills and the minimum experiment height of 0.100 m. This work paves the way for pre-service teachers to use Phyphox to determine the magnitude of gravitational acceleration using three mathematical analyses ranging from the most basic (i.e. taking average) to the most advanced (i.e. commercial plot) in order to assist them in assigning the free-falling experiment to schoolchildren using the most appropriate method of analysis.

Keywords: Phyphox, mobile phone experiment, experimental tools, gravitational acceleration

1. Introduction

The gravitational field strength is simply measured in terms of gravitational acceleration, which is a parabolic function of both displacement and time (Tate, 1968; White et al., 2007). Accurate estimation of its magnitude at various heights is critical for gaining an advantage in transitioning from school (Wang et al., 2017) to military (Chakravarthi, 2011) or astronomy work (Micheli et al., 2018). For example, using a ticker timer to measure the acceleration of gravity is considered to be a typical way in physics classrooms to highlight schoolchildren' potential in scientific data analysis (Fontana, Yeung & Hall, 2020), although measuring gravity with an absolute gravimeter is frequent at either a national or international institution with an accuracy of one in a billion parts (Ménoret et al., 2018). The desire for high-resolution gravity data, technological advancements in device downsizing, and a growing desire for knowledge and information, all have the potential to move the paradigm away from traditional fixed locations and toward mobile monitoring (Widiatmoko, Srigutomo & Kurniasih, 2012). Consequently, to quantify and estimate the magnitude of the gravitational acceleration, mobile sensing and monitoring devices with potential applications such as mobile phones are emerging (Kittiravechote & Sujarittham, 2020; Kuhn, 2014; Kuhn & Vogt, 2013; Pili, Violanda & Ceniza, 2018).

Because a mobile phone operates with a sensor known as micro-electro-mechanical systems, it has been an understandable choice for rapid laboratory equipment to report the strength of gravity (Kuhn & Vogt, 2013). Moreover, several studies use a microphone port or a magnetic field sensor to collect and compute the necessary data, such as the change in sound frequency (Kuhn, 2014) or the period of oscillation (Pili, Violanda & Ceniza, 2018), which leads to the determination of gravitational acceleration. More recently, we have published a paper that used a mobile phone with Phyphox application operating in timer mode and an acoustic stopwatch function to gauge the time of two acoustic events that were started and terminated by two sounds consecutively, and then evaluated the magnitude of the gravitational acceleration (Kittiravechote & Sujarittham, 2020). According to our findings, the experimental result for gravitational strength is determined nearly the standard value of 9.783 m/s^2 provided by the National Institute of Metrology (Thailand). Towards this end, all of these findings emphasize the advantages of using a mobile phone to measure the gravitational acceleration.

In this work, we are interested to address whether there is a difference between novices and experts in determining the acceleration of gravity using a mobile phone with Phyphox application and three mathematical analyses: taking average after formula substitution, manual plot, and commercial plot. To address the question, we utilize the inference statistic t-test to compare the responses of novices and experts: Welch's t-test is used to find the difference between novices and experts, while the paired sample t-test is used to illustrate the difference on three mathematical analyses done by novices. As such mathematical analysis is one of many ways for schoolchildren to interpret scientific data using methods ranging from the most basic (i.e. taking average) to the most advanced (i.e. commercial plot), this study suggested future work to help pre-service teachers in understanding where schoolchildren's difficulties originate during experimentation, resulting in the selection of the most appropriate method of analysis for schoolchildren at various levels.

2. Experiment

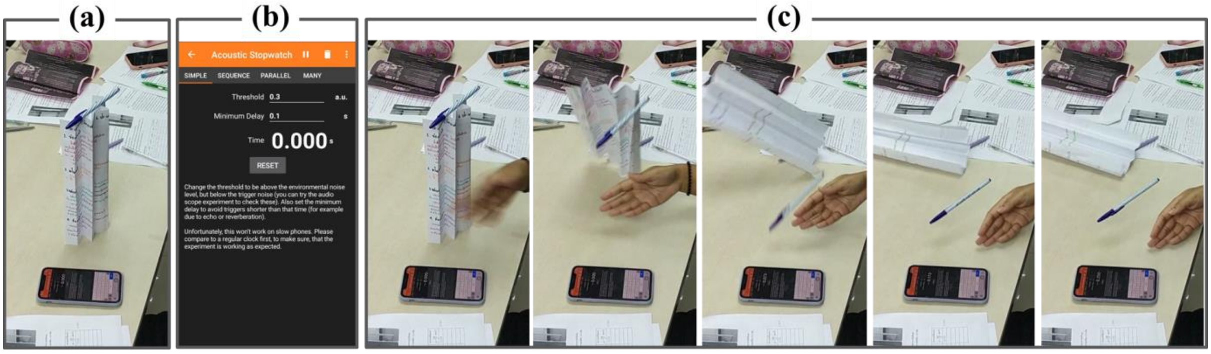
2.1 Mobile Phone Application

This work makes use of Phyphox application (physical phone experiment, version 1.1.7), a free program with at least 30 functions for conducting physics experiments (Staacks et al., 2018). The acoustic stopwatch feature is used as a timer to get the time between two loud acoustic signals. Clicks, beeps, claps, and other sounds are allowed as long as they are louder than ambient noise. The threshold is simply set to 0.3 arbitrary units (from 0 to 1), indicating the level at which the stopwatch is triggered. After this feature is enabled, the stopwatch starts with the first sound that exceeds the threshold and ends with the second sound. To repeat the experiment, simply delete the data and start over. It is important to note that the first sound is brief because a long sound could be misinterpreted as a pause.

2.2 Method

The procedure shown in figure 1 has previously mentioned for conducting the experiment (Kittiravechote & Sujarittham, 2020). In brief, A4 papers taped and folded into a design capable of holding a pencil at a height of approximately 0.100–1.500 meters are used, as illustrated in figure 1 (a). Phyphox application with a timer mode and an acoustic stopwatch feature is used to measure the time spent free-falling, as presented in figure 1 (b). A quick flick of the paper activates Phyphox application and displays the time on the screen, as shown in figure 1 (c). The stopwatch starts and stops timing in response to the first sound of the paper flicking and the second sound of the pencil hitting the floor. After recording and analyzing the results between height in meters and time in seconds, the magnitude of gravitational acceleration is calculated. It should be noted that this experiment is carried out at five different heights, each of which was replicated three times.

The data analysis is conducted by doing as described by Kittiravechote and Sujarittham (2020). In a nutshell, novices and experts are given the task of reporting the averaged free-fall time at each of the five heights, allowing the magnitude of the gravitational acceleration (g) to be determined by three methods: 1) taking the average of acceleration at all five different heights after substituting the averaged time (t) and height (h) into the formula $g = 2h/t^2$, 2) manual plot the linear graph $h = gt^2/2$ when h and t^2 present the data along the y and x axes, respectively, and the twice of its slope becomes the magnitude of gravity, and 3) using a commercial application like Microsoft Excel to map the straight line and estimate the magnitude of gravity (double the slope) (Meyer & Avery, 2009; Ose, 2016).



Threshold 0.3 a.u.

Minimum Delay 0.1 s

Time 0.000 s

Change the threshold to be above the environmental noise level, but below the trigger noise (you can try the audio scope experiment to check these). Also set the minimum delay to avoid triggers shorter than that time (for example due to echo or reverberation).

Unfortunately, this won't work on slow phones. Please compare to a regular clock first, to make sure, that the experiment is working as expected.

Figure 1. Experimental Method. (a) Setting the object's height. (b) Phyphox application with a timer mode and an acoustic stopwatch feature to measure the time spent free-falling. (c) Time series images capture the falling event.

2.3 Participants

35 pre-service teachers (novices) from Program of General Science, Faculty of Education, Bansomdejchaopraya Rajabhat University and 5 physics or general science lecturers (experts) were participated in this work. They all worked independently.

2.4 Statistical Inferences

Welch’s t-test was used to assess the statistical significance of the difference in average changes between novices and experts. Paired sample t-test was used to determine the significance of the difference in average changes between the novices' mathematical analyses.

3. Result

The determination of the magnitude of gravitational acceleration (in m/s^2) obtained by 35 novices and 5 experts are displayed in figure 2. As shown in figure 2 (a) and (b), the mean and standard deviation of the magnitude of the gravitational acceleration for novices, g (taking average) = 9.469 ± 1.057, g (manual plot) = 10.120 ± 1.978, and g (commercial plot) = 9.452 ± 0.897, and for experts, g (taking average) = 9.618 ± 0.315, g (manual plot) = 9.691 ± 0.192, and g (commercial plot) = 9.701 ± 0.223. The results indicate that the mean values of both groups are comparable, e.g., g (taking average): 9.469 for novices vs. 9.618 for experts, g (manual plot): 10.120 for novices vs. 9.691 for experts, and g (commercial plot): 9.452 for novices vs. 9.701 for experts. The difference between the two means g by taking average, manual plot, and commercial plot are 0.149, 0.429, and 0.249 m/s^2, respectively, a marginal difference, suggesting that laboratory studies with novices and experts achieve nearly identical positive effects on the magnitude of gravitational acceleration. The straight lines connecting the three data points are also noted that they were all obtained during the same experiment with the same participant. In addition, the data distributions represented by boxplots in figure 2 (c) for novices and (d) for experts indicate that they are not only free of outliers but also nearly symmetrical, or in other words, the shapes are not overly skewed. Therefore, the possibility of using Welch's t-test to determine the significance of the differences in average changes between novices and experts is now available.

Thanks to the robustness of Welch’s t-test for a comparison of two means with unequal sample sizes (i.e. 5 and 100) and unequal variances (i.e. 1 and 2) under normality as described by Derrick, Toher and White (2016), and thus the degrees of freedom of Welch's t-test are a random variable based on the sample size and variance of each sample, allowing us to perform Welch's t-test for 35 novices and 5 experts. Consequently, the null hypothesis: mean(Novices) – mean(Experts) = 0, along with the alternative hypothesis: mean(Novices) – mean(Experts) ≠ 0 were then defined. Following the calculations, the results in table 1 showed that the t-values are 0.655 with 21 degrees of freedom (df), 1.243 with 37 df, and 1.372 with 27 df for taking average, manual plot, and commercial plot, respectively. According to the t-distribution table (two-tails), all of the obtained t-values for the .10

level are less than the specified values: 0.655 < 1.721 at 21 df, 1.243 < 1.687 at 37 df, and 1.372 < 1.703 at 27 df. As a result, the null hypothesis was accepted in its entirety, while the alternative hypothesis was rejected. In other words, the mean differences between novices and experts were negligible at the 10% level. We suggested that the findings from novices and experts are not different by using a mobile phone with phyphox application to time individual falling objects at each height, together with three mathematical analyses: taking average by substitution of formulas, manually plotting of graphs, and plotting graphs with a program in order to find the gravitational acceleration of the earth at Bangkok.

In addition, paired sample t-test was used to see whether the difference in average changes between the novices' mathematical methods in figure 2 (a) was significant. According to the findings in table 2, there is no significant difference in the magnitude of gravitational acceleration when novices used the methods of taking average after substitution into the formula or commercial plot to evaluate the magnitude as our t-value (0.200) is less than the t-value at .05 level (2.032). When the novices graph on it manually, however, there is a noticeable difference at the .05 level as our t-values (2.244 and 2.051) are higher than the specified ones. As a result, because of the significant difference at the .05 level, there might be a feature of unpredictability in novices' manual plot capabilities.

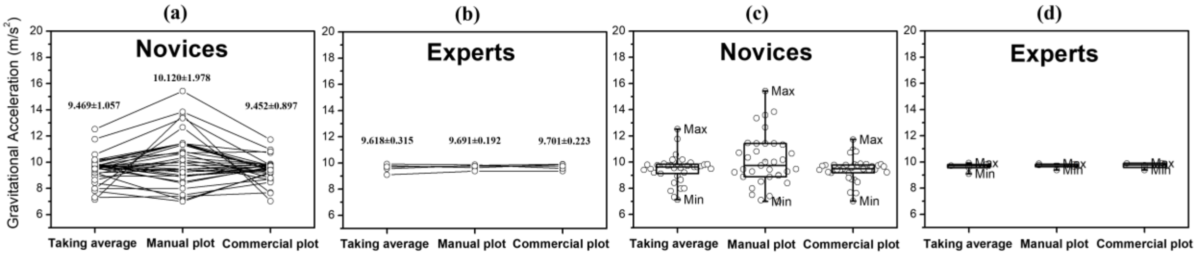


Figure 2. The Determination of the Magnitude of Gravitational Acceleration (in m/s^2). Measures of central tendency and spread (mean ± standard deviation) from novices (a) and experts (b). Nearly symmetric boxplots with no outliers from novices (c) and experts (d) provide the possibility of using Welch's t-test for significant difference.

Table 1. *Welch's T-Test for the Significant Differences in Performance between Novices and Experts*

	Taking average	Manual plot	Commercial plot
Welch's t-test	$t=\dfrac{\lvert 9.469\text{-}9.618\rvert-0}{\sqrt{\dfrac{1.057^2}{35}+\dfrac{0.315^2}{5}}}$ $t=0.655$ $df=21$	$t=\dfrac{\lvert 10.120\text{-}9.691\rvert-0}{\sqrt{\dfrac{1.978^2}{35}+\dfrac{0.192^2}{5}}}$ $t=1.243$ $df=37$	$t=\dfrac{\lvert 9.452\text{-}9.701\rvert-0}{\sqrt{\dfrac{0.897^2}{35}+\dfrac{0.223^2}{5}}}$ $t=1.372$ $df=27$
Noted:	$t=\dfrac{\left\lvert \overline{X}_{Novices}-\overline{X}_{Experts}\right\rvert-\left\lvert \overline{\mu}_{Novices}-\overline{\mu}_{Experts}\right\rvert}{\sqrt{\dfrac{S.D._{Novices}^2}{n_{Novices}}+\dfrac{S.D._{Experts}^2}{n_{Experts}}}}$	$df=\dfrac{\left(\dfrac{S.D._{Novices}^2}{n_{Novices}}+\dfrac{S.D._{Experts}^2}{n_{Experts}}\right)^2}{\dfrac{\left(\dfrac{S.D._{Novices}^2}{n_{Novices}}\right)^2}{n_{Novices}-1}+\dfrac{\left(\dfrac{S.D._{Experts}^2}{n_{Experts}}\right)^2}{n_{Experts}-1}}$	

Table 2. *Paired Sample T-Test for the Significant Differences between the Novices' Mathematical Methods*

Comparison	T-statistic with 34 df	Decision
Taking average vs. manual plot	2.244	Significantly different
Taking average vs. commercial plot	0.200	Not significantly different
Manual plot vs. commercial plot	2.051	Significantly different

Noted: T-statistic with 34 df at .05 level is 2.032

4. Discussion

In attempt to discover the gravitational acceleration of the earth at Bangkok, we have demonstrated how novices and experts use a mobile phone with Phyphox application to time individual falling objects at each height, as well as three mathematical analyses: taking average, manual plot, and commercial plot. At the .10 level, we found that novices and experts performed equally, suggesting that our method of measuring gravitational acceleration is practical and good enough for pre-service teachers to conduct classroom experiments. Further to that, for the novices with the use of three mathematical analyses, we found an insignificant difference between the methods of taking average after formula substitution and commercial plot, but not in the case of manual plot at the .05 level.

To validate the reliability of the magnitude of gravitational acceleration obtained from our experiment, we compare all means to the theoretical value of 9.783 m/s^2 proposed by the National Institute of Metrology (Thailand). All of the means (9.469, 10.120, 9.452, 9.618, 9.691, and 9.701 m/s^2) are comparable to the theoretical value. The maximum error is found to be 3.445 % when novices conduct the manual plot for data analysis. Moreover, the results of the Welch's t-test show that the difference between novices and experts in performing the experiment to determine the earth's gravitational acceleration is insignificant at 90% confidence. As a result, our experimental methods for measuring gravitational acceleration can produce reliable results.

As previously stated, the results of the pair sample t-test show significant differences in novices' data analysis using manual plots compared to taking average after formula substitution or commercial plots at 95 percent confidence interval, implying that there may be a variable that limits novices' manual plotting skills. We hence examine the novices' hand-drawn graphs and summarize the mistakes made by 22 out of 35 novices, as shown in table 3. Not surprisingly, these mistakes are identified as the most common graphical mistakes made by schoolchildren when it comes to slope conceptualization (Hattikudur et al., 2012). These findings suggested that pre-service teachers could identify and recognize faults in manual plot, allowing them to comprehend the problems that schoolchildren face with manual plot as well as the sources of errors or misconceptions.

In addition, we found that when novices performed the experiments at a height of less than 10 cm (about half the width of an A4 paper), some of them showed a misrepresentation of acceleration. They experienced the difficulties of a mobile phone in retrieving the falling time due to the use of a phone that was too late (which could be easily solved by changing the tool to the faster one). This also included Phyphox program's period time problem. However, since the object required 142 milliseconds to reach the ground at a height of 10.0 centimeters, Phyphox developer recommended that the acoustic stopwatch wait at least 100 milliseconds before accepting a second signal (AachenUniversity, November 2017). The results indicate limitations in our experiments, especially the need for a suitable height of greater than 0.100 m above the ground.

Table 3. *A Summary of the Mistakes Found in 22 of 35 Novices*

Description of mistakes	Novices
Incorrect scale. On the x-axis, y-axis, or both x and y-axes, the scale is incorrect. Novices should represent the same value on each axis, for example, if each division is set to 0.005 at the start, then the numbers along the axis should be 0.005, 0.010, 0.015, and so on. As previously discussed, this provides them to estimate the magnitude of gravity from the slope of best fit passing through the origin.	3
Unbalanced line of best fit. (1) Data points above and below the line of best fit appear to be less evenly distributed. (2) Draw a line that best describes the identified trend as a connection of data points, rather than a straight line. Novices should draw the linear line that best applies to the majority of the data points and should be less concerned with data points that differ from the majority. (3) Line of best fit does not pass through the origin.	22
Calculate slope from two specific data points. Using two raw data points to determine the graph's slope. To calculate the slope of a graph, novices could choose any two points that lie on the line of best fit.	2
Missing unit. On the x and y axes, the variables and units of measurement are missing. They can be identified by novices.	5

5. Conclusion

We have demonstrated the differences in how novices and experts conduct the experiment to determine the magnitude of the gravitational acceleration using a mobile phone with Phyphox application, as well as the data analysis with three mathematical methods: taking average after formula substitution, manual plot, and commercial plot. Not only do the responses of novices and experts match the theoretical value, but we also present a statistically negligible difference in their performance with a 90% confidence interval. Accordingly, our method of determining gravitational acceleration is reliable, making it possible for pre-service teachers doing classroom experiments with schoolchildren. Noted that the responses from novices after using the manual plot for data analysis differ significantly from those obtained by taking average and commercial plot with 95% confidence interval due to some limitations related to novice plotting skills and the minimum experiment height of 0.100 m.

Acknowledgements

Research and Development Institute, Bansomdejchaopraya Rajabhat University is acknowledged.

References

AachenUniversity, R. W. T. H. (November 2017). *Experiment: Acoustic Stopwatch*. https://phyphox.org/wiki/index.php/Experiment:_Acoustic_Stopwatch.

Chakravarthi, V. (2011). *Encyclopedia of Solid Earth Geophysics*. Dordrecht: Springer Netherlands.

Derrick, B., Toher, D., & White, P. (2016). Why Welch's test is Type I error robust. *Quantitative Methods for Psychology, 12*(1), 30-38.

Fontana, E., Yeung, C., & Hall, J. C. (2020). Determining the acceleration due to gravity and friction using the ticker tape timer method. *The Physics Teacher*, *58*(5), 338-339.

Hattikudur, S., Prather, R. W., Asquith, P., Alibali, M., Knuth, E. J., & Nathan, M. (2012). Constructing graphical representations: Middle schoolers' intuitions and developing knowledge about slope and y-intercept. *School Science and Mathematics*, *112*(4), 230-240.

Kittiravechote, A., & Sujarittham, T. (2020). Measuring the acceleration of gravity using a smartphone, A4-papers, and a pencil. *International Journal of Advanced Science and Technology*, *29*(7s), 884-889.

Kuhn, J. (2014). Relevant information about using a mobile phone acceleration sensor in physics experiments. *American Journal of Physics*, *82*(2), 94-94.

Kuhn, J., & Vogt, P. (2013). Smartphones as experimental tools: Different methods to determine the gravitational acceleration in classroom physics by using everyday devices. *European J of Physics Education*, *4*(1), 16-27.

Ménoret, V., Vermeulen, P., Le Moigne, N., Bonvalot, S., Bouyer, P., Landragin, A. et al. (2018). Gravity measurements below 10−9 g with a transportable absolute quantum gravimeter. *Scientific Reports*, *8*(1), 12300.

Meyer, D. Z., & Avery, L. M. (2009). Excel as a qualitative data analysis tool. *Field Methods*, *21*(1), 91-112.

Micheli, M., Farnocchia, D., Meech, K. J., Buie, M. W., Hainaut, O. R., Prialnik, D. et al. (2018). Non-gravitational acceleration in the trajectory of 1I/2017 U1 ('Oumuamua). *Nature*, *559*(7713), 223-226.

Ose, S. O. (2016). Using excel and word to structure qualitative data. *Journal of Applied Social Sciences*, *10*(2), 147-162.

Pili, U., Violanda, R., & Ceniza, C. (2018). Measurement of g using a magnetic pendulum and a smartphone magnetometer. *The Physics Teacher*, *56*(4), 258-259.

Staacks, S., Hütz, S., Heinke, H., & Stampfer, C. (2018). Advanced tools for smartphone-based experiments: phyphox. *Physics Education*, *53*(4), 045009.

Tate, D. R. (1968). Acceleration due to gravity at the National Bureau of Standards. *Journal of Research of the National Bureau of Standards, Section C: Engineering and Instrumentation*, *72C*(1), 1-20.

Wang, Q., Wang, C., Xiao, Y., Schulte, J., & Shi, Q. (2017). A new method of measuring gravitational acceleration in an undergraduate laboratory program. *European Journal of Physics*, *39*(1), 015701.

White, J. A., Medina, A., Román, F. L., & Velasco, S. (2007). A measurement of g listening to falling balls. *The Physics Teacher*, *45*(3), 175-177.

Widiatmoko, E., Srigutomo, W., & Kurniasih, N. (2012). Measurement of gravitational acceleration using a computer microphone port. *Physics Education*, *47*(6), 709-714.

Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Effects of Virtual Reality on Students' Creative Thinking during a Brainstorming Session

Mondheera PITUXCOOSUVARN[a*], Victoria ABOU-KHALIL[b], Hiroaki OGATA[b] & Yohei MURAKAMI[a]
[a]*Faculty of Information Science and Engineering, Ritsumeikan University, Japan*
[b]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*mond-p@fc.ritsumei.ac.jp

Abstract: Brainstorming is a well-known technique for fostering student creativity. Due to the COVID-19 pandemic, brainstorming sessions were recently held online, using web-based tools and video calls. Virtual Reality (VR) can also be an alternative for brainstorming sessions. However, there is currently limited research on brainstorming sessions using VR, and its impact on students' creative thinking is still unknown. We conducted a preliminary study that compares brainstorming in VR and brainstorming with a web-based online whiteboard to study how each communication method affects students' creative thinking. Given that students had the same amount of time for both VR and web sessions, the results reveal that there is no significant difference in the quality or quantity of ideas. Even though participants said VR was difficult to use, the results of VR and web sessions were similar. We believe that if the students become familiar with VR, they will be able to develop more ideas in the virtual space.

Keywords: Brainstorming, online learning, virtual reality, online whiteboard, collaborative learning

1. Introduction

During the COVID-19 pandemic, online activities became common education methods. There are various ways to deliver lessons to the students online including live video streaming, asynchronous video broadcasting, or synchronous online classes. In classes that are interactive and collaborative, the communication medium plays a fairly big role (e.g., design classes; policy studies classes; project-based learning classes). Brainstorming is a standard activity that is often used to foster student creativity in face-to-face classes. When the brainstorming is conducted online, online whiteboard tools are generally used, e.g., Google's Jamboard, Miro board, etc. These tools allow the students to share a whiteboard online with their peers and brainstorm together on ideas. Regularly used functions of the online whiteboard for brainstorming are typing messages, drawing, putting on sticky notes, rearranging the notes and messages. Online whiteboards enable online brainstorming and meeting and are usually used together with video call applications, while some of the online whiteboards have the video call function included as a part of their system. Besides web-based whiteboards, Virtual Reality (VR) could be an alternative workspace for distant creative sessions. In the last few years, VR has exploded in popularity as the headsets have been available for general users. A VR headset is a head-mounted device that immerses the user in virtual reality. Using VR, it is possible for students to also interact and have a brainstorming session online.

There exist various studies on VR's effect on creativity (Lin, Wang, Kuo, & Luo, 2017), however, studies focusing on comparing VR brainstorming sessions and brainstorming sessions using an online whiteboard has not been conducted. In this paper, we study the difference between online whiteboard brainstorming sessions and VR sessions in terms of quantity of ideas, quality of ideas, cognitive load, and satisfaction.

2. Background

2.1 Brainstorming

Brainstorming is known as an individual or group method for creating ideas, increasing creativity, and finding solutions (Wilson, 2013). Based on Osborn (1953), there are four rules for brainstorming. First and foremost, no evaluation should take place during the session, no matter how absurd the ideas may appear. Second, the team should generate as many ideas as possible. Third, wild and crazy ideas are welcome. And the last, creating new ideas on top of each other's ideas is important. One of the well-known brainstorming practices is to write ideas on sticky notes and collaboratively discuss and group them (Kumar, 2012).

2.2 VR and Creativity

Thornhill-Miller and Dupont (2016) studied if VR enhances creativity and innovation, under-recognized potential among converging technologies. They reported that VR provides a cost-effective means of implementing and optimizing nearly all conventional creativity enhancement techniques, while also providing powerful new possibilities that are not available through traditional means. In 2017, Lin, Wang, Kuo, and Luo (2017) studied the effect of virtual reality 3D exploratory education. Their result showed that students' creativity and leadership in exploratory education produced the highest creativity with VR.

3. Methods

3.1 Experiment Design

We conducted a repeated measures study considering the system used for brainstorming as the independent variable. The dependent variables are the quality and quantities of the generated ideas, the perceived ease of use of the system as well as its perceived usefulness. The study participants were three graduate students at a university in Japan. All students were enrolled in the same program in the graduate school of informatics and had experienced brainstorming one to five times before the experiment.

The study was conducted over two phases as shown in Fig. 1. As a first step, we introduced the rules of brainstorming and conducted a brainstorming session with the participants using paper and pen to reduce the ordering effect. During the first phase, we first asked the participants to use the virtual reality headset for twenty minutes to become familiar with it. After that, participants were asked to use the VR headset and provide ideas to the problem used by Hender et al.(2002): "A restaurant located next to campus is losing customers. What can the restaurant do to retain its customers?" The participants used the Spatial application on the Oculus Quest 2 headset for the brainstorming session conducted using VR. Spatial allows participants to meet in a meeting room, scribble on a whiteboard and post sticky notes on the board as shown in figure 2. Phase 2 consisted of a web-based brainstorming session using Zoom for the meeting and Google Jamboard for the ideation session. For the web-based brainstorming session, the participants were assigned a modified version of Hender et al.(2002) problem: "The university library campus is visited less. What can the library do to attract more students?"

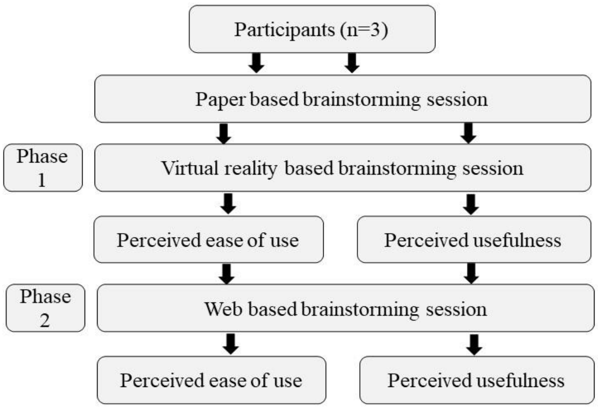


Figure 1. Experimental Procedure.

Both the VR and web-based brainstorming sessions were structured similarly. First, the facilitator presented the problem and asked the participants to generate ideas and write them on sticky notes for a period of ten minutes. Once completed, the participants were asked to share their ideas and explain them. The participants were then given five minutes to generate additional ideas and share them. After that, the participants were asked to group similar ideas together, name the groups, and choose their final solution by discussion.

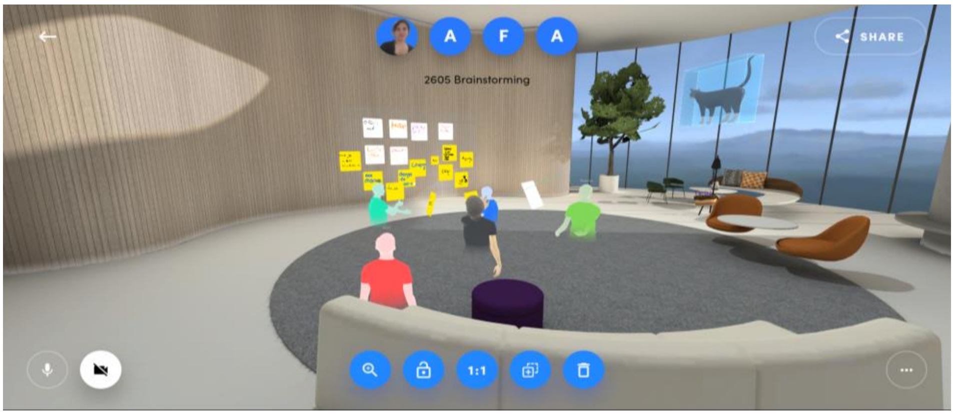


Figure 2. Overview of the Brainstorming Session using VR.

3.2 Dependent Variables

We defined four dependent variables for this experiment as follows:

- *Quantity of ideas:* The quantity of ideas was determined by counting the number of ideas generated by each participant and summed.
- *Creativity of ideas:* The creativity of the idea was measured in terms of the originality of the idea. Two raters coded all ideas for originality. A 5-point scale was used (1 as not original, 5 as highly feasible). Examples of a highly original and a highly unoriginal idea, respectively, are: “Create a book swapping campaign in the library” and "install comfortable chairs in the library" and “Teach courses in smaller groups.” The rater’s agreement was measured by considering that the raters are in agreement if the ratings do not have more than one point of difference (Diehl & Stroebe, 1987). An inter-rater Agreement was present in 96.4% of the cases. As the inter-rater agreement is high, we used the scores of the first rater for our analysis. We calculated the mean originality for the ideas generated using VR and the ideas generated using the web.
- *Ease of use:* Perceived ease of use was measured to assess the cognitive load of each system. Perceived ease of use was measured using the instrument developed by Sambamurthy & Chin(1994). The ease of use was measured using 5-point Likert-scale questions where 1= strongly disagree and 5=strongly agree. The responses for the questions were summed to calculate the perceived ease of use (Hender et al., 2002).

- *Usefulness:* Perceived usefulness was measured to assess the usefulness of each system to conduct brainstorming sessions. Perceived usefulness was measured using the and adaptation of the instrument developed by Sambamurthy & Chin(1994). The usefulness was measured using 5-point Likert-scale questions where 1= strongly disagree and 5=strongly agree. The responses for the questions were summed to calculate the perceived usefulness.

4. Results

A t-test was used to analyze the differences in the number of ideas, the creativity of ideas, perceived ease of use, and perceived usefulness between a VR-based brainstorming session and a web-based brainstorming session.

As displayed in Table 1, there was no significant difference in the number of ideas generated using VR compared to the use of the web application. There was also no significant difference in the quality of ideas generated using VR compared to the use of the web application.

The ease of use of web-based brainstorming sessions is significantly greater than the ease of use of VR-based brainstorming sessions.

The usefulness of web-based brainstorming sessions is significantly greater than the usefulness of VR-based brainstorming sessions.

Table 1. *Quantity and quality of ideas, ease of use, and usefulness of brainstorming using VR and web*

	Virtual Reality Session			Online Whiteboard Session			t	Cohen's D
	N	Mean	SD	N	Mean	SD		
Quantity	3	9.66	0.47	3	9	2.1	0.42	0.42
Quality	29	2.89	0.92	29	2.68	0.64	0.97	0.26
Ease of use	3	15.66	1.69	3	18.33	1.24	-1.78*	1.8
Usefulness	3	6.66	2.05	3	11	0.81	-2.77**	2.78
*p<0.1, **p<0.05								

5. Discussion

According to the results of the experiment, the students seem to prefer the online whiteboard over the VR. They also mentioned the challenges they had using VR headsets while they mentioned being comfortable with the online whiteboard since they are familiar with web technologies. Even though using VR students reported facing difficulties with the VR and taking longer time to write and attach the sticky notes, they still could generate as many ideas as on the web-based session. When VR becomes more a normal way of communication, in other words, if the students become more familiar with VR, it is highly possible that the students will contribute more ideas in the virtual realm.

There are limitations in this study as the students are more familiar with using computers and the web than VR headsets. The results would be more accurate if the students are familiar with both technologies. In the near future, we plan to conduct several experiments with different groups of participants and with different levels of VR experiences to confirm the validity of this experiment and our assumption about the number of ideas that might be increased when the students are more comfortable with VR.

6. Conclusion

This paper presents a preliminary study in which we compared brainstorming in the VR to brainstorming on a web-based online whiteboard to determine how each communication style influences students' creative thinking. We asked graduate students to brainstorm in the VR using VR headsets and brainstorm on the web using an online whiteboard and video call. The students reported that the ease of use in VR is lower than in the web-based session. However, they could produce the similar number of ideas with similar quality.

Acknowledgements

This research was partially supported by a Grant-in-Aid for Scientific Research (A) (17H00759, 2017-2020), a Grant-in-Aid for Scientific Research (B) (21H03561,2021-2024) and a Grant-in-Aid for Early-Career Scientists (21K17794,2021-2024) from the Japan Society for the Promotion of Sciences (JSPS). A part of this paper was also supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (S) 16H06304 and NEDO JPNP20006 and JPNP18013.

References

Diehl, M., & Stroebe, W. (1987). Productivity loss in brainstorming groups: Toward the solution of a riddle. Journal of Personality and Social Psychology, 53(3), 497.

Hender, J. M., Dean, D. L., Rodgers, T. L., & Nunamaker Jr, J. F. (2002). An examination of the impact of stimuli type and GSS structure on creativity: Brainstorming versus non-brainstorming techniques in a GSS environment. *Journal of Management Information Systems*, *18*(4), 59–85.

Kumar, V. (2012). 101 design methods: A structured approach for driving innovation in your organization. John Wiley & Sons.

Lin, M. T. Y., Wang, J. S., Kuo, H. M., & Luo, Y. (2017). A study on the effect of virtual reality 3D exploratory education on students' creativity and leadership. Eurasia Journal of Mathematics, Science and Technology Education, 13(7), 3151-3161.

Osborn, A. F. (1953). Applied imagination.

Sambamurthy, V., & Chin, W. W. (1994). The effects of group attitudes toward alternative GDSS designs on the decision-making performance of computer-supported groups. Decision Sciences, 25(2), 215–2

Thornhill-Miller, B., & Dupont, J. M. (2016). Virtual reality and the enhancement of creativity and innovation: Under recognized potential among converging technologies?. Journal of Cognitive Education and Psychology, 15(1), 102-121.

Wilson, C. (2013). Brainstorming and beyond: a user-centered design method. Newnes.

Karyotype: An Interactive Learning Environment for Reasoning and Sense Making in Genetics through a Case-based Approach

Sunita RASTE[a*], Anurag DEEP[a] & Sahana MURTHY[a]

[a]*Indian Institute of Technology Bombay, India*

*sunita_20@iitb.ac.in

Abstract: Reasoning and sense-making has been proved to be difficult for learners. Connecting genetic disorders with their underlying chromosomal aberrations requires reasoning and sense-making based on the clinical symptoms. In traditional instructional methods, learners encounter the required concepts and scientific processes in different grades and across different settings – theory classes, labs, tutorials, and most tests assess learners' ability to remember the facts. This leads to a focus on vocabulary and memorization, rather than on scientific reasoning and making connections across different concepts and processes. To address this gap, we have developed *Karyotype*: a web-browser based learning environment where learners assume the role of a geneticist and solve the cases of chromosomal disorders. The learning environment is based on a case-based reasoning approach where the learner acts as a problem solver. S/he is given a set of symptoms and is asked to explain them and suggest a diagnosis. In this paper, we present the theoretical basis and design of *Karyotype*. We report an exploratory study in which we investigated learning of problem-solving skills and perceptions of usability and usefulness of techno-pedagogical features of *Karyotype*. The results indicate that learning activities and scaffolds in *Karyotype* are helpful for learning reasoning and sense-making skills.

Keywords: Interactive content, chromosomal disorders, reasoning, sense-making, case-based approach

1. Introduction

A genetic disorder is a disease caused in whole or in part by a change in the DNA sequence away from the normal sequence. One such type is chromosomal disorders, caused due to abnormalities such as a change in the number or structure of entire chromosomes, or a specific part of the chromosome carrying a particular gene. Students pursuing bioscience majors encounter these concepts to varying levels of breadth and depth in their undergraduate and postgraduate curriculum. They learn about genetic mutations that cause disorders, their symptoms, and their effects as well as the diagnostic methods and treatments. However, often they learn these topics in the form of disjoint facts, which they perceive as a memorization activity. Students fail to make the connection between the defects at the chromosomal level and their corresponding effect on the features displayed by an individual. Making these connections requires reasoning in terms of physiological states, complaints, and symptoms. Typical curricula cover the concepts and procedures required for diagnosis in silos in theory classes, practical labs, and tutorial sessions. What gets missed is the emphasis to facilitate learners through the reasoning process and making explicit connections across the concepts.

An important process in diagnosing genetic disorders is karyotyping. The karyotyping activity has been traditionally conducted as a pen and paper exercise where learners are provided with a print-out of the chromosome spread. Learners are expected to arrange the chromosomes in a particular order as per their size, the position of the centromere, and the length of p and q arms of a chromosome. This is done by matching the homologous pairs to complete the karyotype which is a map of human chromosomes. After completion of the karyotype, the anomaly can be detected as any additional/missing chromosome. An extension of the traditional pen and paper activity includes karyotyping kits developed by the Science Source (The Science Source SE, 2021), Thermo Fisher

Scientific (Thermo Fisher Scientific, 2021), and Carolina (Carolina, 2021). While these products are in the form of tangible objects available offline, there are a few efforts made to bring the interactivity elements via online simulations (Labster Cytogenetic, 2021), virtual labs (Virtual cell biology lab, 2021), and web-based approaches (The Biology Project, 2021; Learn Genetics, 2021; Training Karyotypes, 2021). These systems provide stand-alone activities that make use of interactivity for the arrangement of chromosomes to complete a karyotype. However, this approach does not focus on learners making connections between the cause and effect associated with an abnormal chromosomal condition. Instructional strategies combined with interactive visual teaching resources would aid students in obtaining a large amount of knowledge and remembering more ideas (Riyanto, Amin, Suwono & Lestari 2020). Few systems provide example cases for the learners before they move on to the karyotyping activity.

There is a need for a learning environment that can provide hands-on experience of the complete process followed for the diagnosis of disorders. Technology-enhanced learning environments can be used for facilitating case-based reasoning by providing overall structure to the interactive learning activities, immediate feedback, scaffolds in form of reflective and evaluative question prompts, and so on (Deep, Murthy & Bhat, 2020). We propose an online learning environment *Karyotype* that provides learners with a series of interactive learning activities situated in the context of specific cases of disorders wherein learners diagnose the disorder based on the underlying genetic condition, while reasoning through different phases of inquiry.

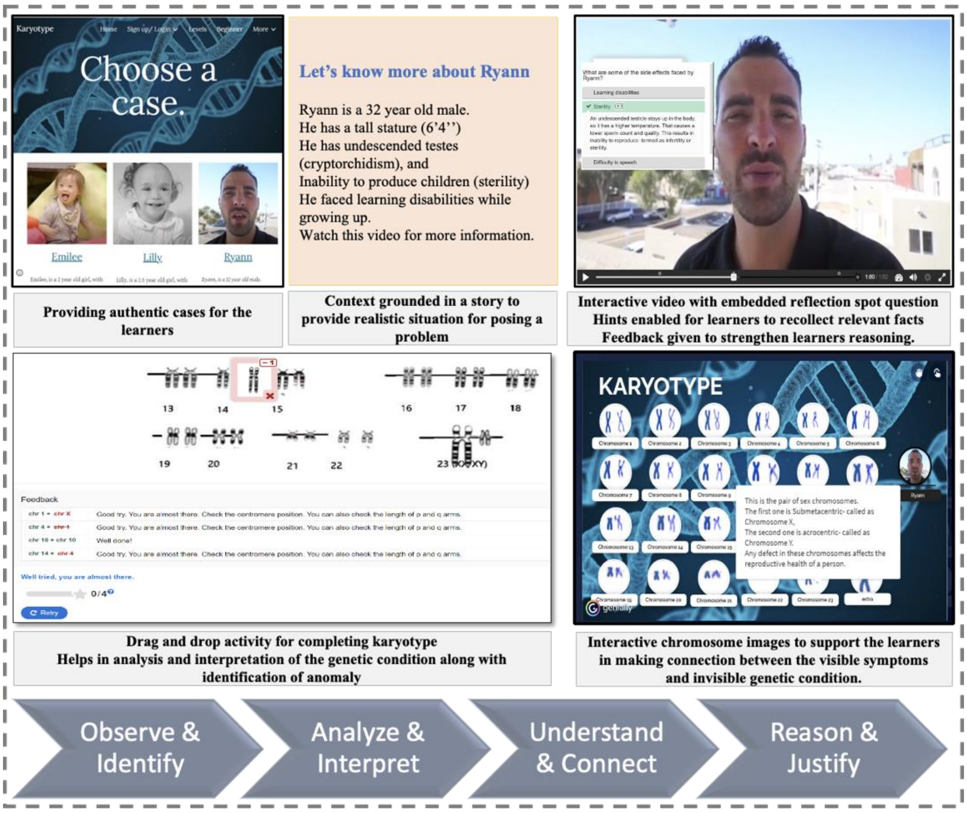
In this paper, we describe the theoretical basis and design of *Karyotype*, and report a pilot study with 5 bioscience majors' learners in the context of chromosomal disorders. The two research objectives of this study are to examine the effectiveness of the scaffolds in the system to help learners in reasoning and sense making, and to understand learners' perception about learning activities.

2. Theoretical Basis

Sense-making about a phenomenon involves collecting observations, analysing the data and constructing interpretations. Sense-making and reasoning are closely associated in order to strengthen the interpretations. Stories are the oldest and most natural form of sense making as they allow explanation and interpretation. It helps to develop empathy towards the central character of the story, which acts as a way to understand a situation better (Herreid, 2007; Jonassen and Hernandez-Serrano, 2002; Centre for teaching and learning 2021). Exposing the learners to different cases or stories while solving problems, provides them an opportunity to reflect in action (Jonassen and Hernandez-Serrano, 2002). Combining the two together is the cased-based learning approach which presents stories in the form of cases to be solved. Case-Based Reasoning (CBR) proposes a method in which students learn by doing problem solving and other activities that pique their interest and allow them to apply what they've learned in a way that provides immediate feedback (Kolodner et al, 2003). CBR has been used in medicine for diagnosis as the methodology of CBR systems closely resembles the thought processes of a physician. This reasoning includes cognitive activities similar to sense making, like gathering information, recognition of patterns, solving problems, and decision making (Choudhury & Ara, 2016). In CBR, the remembered cases are used as a means of efficient problem-solving. Here elicitation becomes a task of gathering case histories and implementation is reduced to identifying significant features that describe a case (Watson and Marrir, 1994).

3. Design and Development of *Karyotype*

Karyotype is an online, self-paced, task-oriented learning environment. The target learners of *Karyotype* are biology majors' students. Educators propagating basic science can make use of *Karyotype* as an instructional system. Additionally, *Karyotype* can cater to learners across the disciplines and age groups, who are interested in acquiring knowledge about the genetic material and its impact on our lives in general. The learning environment has a series of learning activities that helps the learners to have a complete walkthrough of the process followed for clinical diagnosis of chromosomal abnormalities. The system makes use of interactive elements to enhance learners' engagement with the content and provides scaffolds for the learners to help them make progress in the learning activities that are based on various phases of an inquiry cycle which guide learners to achieve the intended learning outcomes. An interactive video narrative presented to the learners in the form of a case history of a patient (Fig 1, top left), acts as an anchor providing an authentic and engaging narration to help learners understand and empathize with the patients' story. Learners have the autonomy to choose any case of their interest based on the case briefing provided. Reflection spots within the video (Fig 1, top right) are used as scaffolds to help in the better understanding of the problem context and identification of the symptoms associated with a disorder. Hints are provided to help learners choose appropriate explanations from the given set of options. Constructive feedback is given to scaffold the learners in concluding the reflection activity. Drag and drop activity along with hints and feedback (Fig 1, bottom left) to move the chromosomes in order to complete a karyotype, helps the learners in analysing and interpreting the underlying genetic condition. This allows identifying the anomaly in a given karyotype. The help and information prompts (Fig 1, bottom right) associated with interactive chromosome images are used to scaffold the learners to make connections between the visible symptoms and invisible genetic conditions. Interactive images act as redundant scaffolds providing more opportunities for learners to access additional information as and when required. Learners are supposed to prepare a diagnosis report based on their observations and the reference material provided. A look-up table is provided to scaffold collating observations and results from previous phases of inquiry for making a final prediction about the disorder. The report needs to have reasons and justifications for the choice made by a learner. The purpose of this is to help improve the reasoning skills of learners (Kolodner et al, 2003). Using the correct set of reasoning would help in making a correct diagnosis, thereby aiding the problem-solving process.



Let's know more about Ryann

Ryann is a 32 year old male.
He has a tall stature (6'4'')
He has undescended testes (cryptorchidism), and
Inability to produce children (sterility)
He faced learning disabilities while growing up.
Watch this video for more information.

What are some of the side effects faced by Ryann?
Learning disabilities
✓ Sterility
An undescended testicle stays up in the body, so it has a higher temperature. That causes a lower sperm count and quality. This results in inability to reproduce- termed as infertility or sterility.
Difficulty in speech

Providing authentic cases for the learners

Context grounded in a story to provide realistic situation for posing a problem

Interactive video with embedded reflection spot question
Hints enabled for learners to recollect relevant facts
Feedback given to strengthen learners reasoning.

Feedback
chr 1 → chr X Good try. You are almost there. Check the centromere position. You can also check the length of p and q arms.
chr 4 → ~~chr 4~~ Good try. You are almost there. Check the centromere position. You can also check the length of p and q arms.
chr 10 → chr 10 Well done!
chr 14 → chr 4 Good try. You are almost there. Check the centromere position. You can also check the length of p and q arms.

Well tried, you are almost there.
0/4
Retry

This is the pair of sex chromosomes.
The first one is Submetacentric- called as Chromosome X,
The second one is acrocentric- called as Chromosome Y.
Any defect in these chromosomes affects the reproductive health of a person.

Drag and drop activity for completing karyotype
Helps in analysis and interpretation of the genetic condition along with identification of anomaly

Interactive chromosome images to support the learners in making connection between the visible symptoms and invisible genetic condition.

Figure 1. Learning Activities in Karyotype.

A few examples of chromosomal disorders included in *Karyotype* are Turner's syndrome, Larsen syndrome, and Jacobsen syndromes. The user interface of *Karyotype* is designed and implemented with Google sites, which is an open-source toolkit. The learning activities of *Karyotype* have been designed in H5P and genially. H5P is an open-source, HTML5 toolkit to develop interactive content. *Karyotype* can be accessed through a standard web browser using any device.

4. Study Design

The purpose of this pilot study was to understand the role of the pedagogical design features of *Karyotype* through participants' experience, performance and perceptions, and inform redesign. The participants of this study were 5 bioscience majors' students from one of the colleges in Kerala, India. In this study, we chose chromosomal disorders as the context covered in the learning material. Problem-solving in this topic requires the students to understand the context of the problem, make relevant observations regarding the case, perform basic tests for diagnosis and come up with a reasoning and justification for a possible diagnosis associated with the given task.

This study was conducted as a part of a workshop for bio-science learners. It was conducted online in the presence of the instructor in a supervised setting. Google Meet was used as the online video conferencing platform for the synchronous session and post workshop interviews. The *Karyotype* learning environment was accessed as a Google Site using the standard web browser for capturing data of the learning gains on case-based reasoning and their perception of scaffolds present in the LE. The study had five steps - Registration and self-perception survey, Pre-Test, Interaction with the LE, Post-test and post workshop interviews. The registration form recorded participants personal and academic information along with their confidence (on a likert scale of 1-5, ranging from low to high) regarding the knowledge of chromosomal abnormalities, preparing a karyotype, analyzing and interpreting a given karyotype, and, reasoning about the causes and symptoms associated with a chromosomal disorder.

There are 4 data sources in the study which includes Pre-test, reports of *Karyotype* learning activities, report from post-test, and interviews. Reports of learning activities include notes and observations made by learners and their diagnosis. The pre-test consisted of two questions to understand how learners make sense of the given problem and the role of providing a context. Each participant solved 3 cases during their interaction with *Karyotype.* After going through a series of learning activities (Fig.1), students generated diagnosis reports which were to understand the role of the scaffolds in the diagnosis process. In the post test, these scaffolds were withdrawn. Participants' responses from the artefacts submitted in the form of a diagnosis report were analysed on three criteria related to sense making: C1 – Making explicit and relevant clinical observations, C2 – Connecting the observable clinical symptoms with the chromosomal condition and C3– Explaining how chromosomal changes lead to clinical symptoms. A 3-point rubric (0-missing, 1-needs work, 2- adequate) was used to analyse the data from pre-test, final diagnosis report and post-test. The semi-structured stimulated recall interviews were conducted to get insights regarding participants' perception about the techno-pedagogical design and learning activities in their process of diagnosis and sense making. Participants were probed about their experience of interacting with *Karyotype* and how they made use of the learning activities to solve the given case. 30 instances of participants' use of case-based learning strategies and *Karyotype* design elements were identified.

5. Findings

5.1 Performance, Usability and Usefulness of Karyotype

Table 1 represents the average performance of participants for the three criteria in the pre-test, within the activities of the intervention and in the post-test.

Table 1. *Participants (P) average scores*

P	Overall			C1			C2			C3		
	Pre	Int.	Post	Pre	Int.	Post	Pre	Int.	Post	Pre	Int.	Post
S1	0.66	1.66	1	0	2	0	1	2	2	1	1	1
S2	1	1.33	1.33	1	2	2	1	1	1	1	1	1

S3	0.66	1.66	1.66	0	2	2	1	2	2	1	1	1
S4	1.33	1.66	1.33	2	2	1	1	2	2	1	1	1
S5	1	1.33	NA*	1	2	NA*	1	1	NA*	1	1	NA*

*Did not attempt due to technical reasons

The majority of learners were found to make explicit and relevant clinical observations. A sample observation by a participant is as follows: *"Emile is suffering from a genetic disorder. The symptoms shown by her are a sudden drop in blood sugar level so she is admitted in NICU and fed through nose. She has been given psychological support to avoid mental breakdown."* However, S1 missed noting down the observations in the post test which could be due to the absence of scaffolds in the post test. Participants found the interactive videos with reflection spots to be useful for making clinical observations. S2: "*There are many questions in between the video which helps to summarize the symptoms of the syndrome. It was very helpful.*" It was found that while making notes and observations about a case, participants made an emotional connection and expressed empathy towards the patient. S4: *"Nora's parents were disappointed when the doctors say there is no hope for the child and if she was born, they can't say how long she will live. When she was born, they were very happy to see her. Now they are very happy about her performance. She can smile, sing. And her father was disappointed about her cleft lips."*

Learners' performance in connecting the observable symptoms with the chromosomal condition (Table 1) improved from "needs work" level in pre-test to "adequate" in learning activities and post-test. This performance remained consistent in the post test after removal of scaffolds suggesting that learners can become a more efficient reasoner by remembering old solutions and adapting them rather than having to derive answers from scratch each time (Kolodner, 1992). *S3: "Hints are always useful. Okay. But towards the end, it's not much difficult for me."* This is a key feature of the case-based reasoning approach considered as a methodology for both reasoning and learning. Participants found the look-up table useful for moving towards the correct diagnosis by matching the symptoms and underlying chromosomal condition. *S1: "Lookup table was required, because while considering this patient's symptoms, we have some doubts that it may be this. So, by looking at this table, we can confirm it. It is this disease." "By seeing the video, we can understand their phenotypic characteristics. And then by the lookup table, we can understand the genotype."* However, the performance in explaining how chromosomal changes lead to clinical symptoms (Table 1) remained constant with no significant increase or decrease suggesting that spending more time with the LE and exposure to more cases might be required.

6. Discussion and Future Work

The techno-pedagogical design elements used in the LE as per the guidelines about the role of scaffolds in software assisted learning (Quintana et al, 2004), when mapped with their intended purpose with respect to the learning outcomes for learners were found to be consistent with the findings from the interviews and artefacts. Making clinical observations and noticing what is important about scientific situations requires substantial conceptual domain-specific knowledge, which novice learners may lack. It's beneficial to incorporate expert input to assist learners in making connections to things they are familiar with while also challenging them to comprehend new phenomena. Participants' responses about the interactive videos and hints/feedback provided during the drag and drop activity, suggest that having such visual representation helps learners understand the problem context better, aid observations and problem solving. Engaging in reflective self-assessment can increase learners' understanding about the content and the inquiry process. It is observed that sometimes learners conceive opportunities for articulation and reflection as mere blank fields to be filled in and miss out on productive reflection. Scaffolds can help by focusing learners' attention on making strategic decisions they might otherwise avoid. Interactive chromosome images provide visual representations that may be examined to expose data's underlying qualities. Look-up table helps engage in sense-making practices while reflecting on disciplinary strategies. The report template provides a mechanism for recording observations, findings, or ideas during the diagnosis process. It also helps the learners to express their thinking in ways that highlight important disciplinary ideas. Our next development work includes analysis of the users' feedback to modify and implement additional functionalities in the learning environment, and conducting further user studies to understand how learners navigate through the process of reasoning and sense-making.

Acknowledgements

We thank Next Education Labs, Educational Technology, IIT Bombay for the support provided in developing *Karyotype.* We thank all the people whose stories and videos were used for creating the learning activities. We thank Prof. Praveen Prasad and our participants from SN College, Nattika for helping us with this research study and Mohd. Saif for the beta testing of the system.

References

Carolina. Magnetic karyotype layout board. (2021). *Retrieved on May 31, 2021.* https://www.carolina.com/human-genetics/karyotyping-with-magnetic-chromosomes-kit/FAM_173837 .pr

Centre for teaching and learning (2021), Queens University, Case based learning. *Retrieved on May 31, 2021.* https://www.queensu.ca/ctl/teaching-support/instructional-strategies/case-based-learning#:~:text=Using %20a%20case%2Dbased%20approach,group%20to%20examine%20the%20case

Choudhury, N., & Ara, S. (2016). A Survey on Case-based Reasoning in Medicine. *International Journal of Advanced Computer Science and Applications*, *7*(8), 136–144.

Chris Quintana, Brian J. Reiser, Elizabeth A. Davis, Joseph Krajcik, Eric Fretz, Ravit Golan Duncan, Eleni Kyza, Daniel Edelson & Elliot Soloway (2004) A Scaffolding Design Framework for Software to Support Science Inquiry, Journal of the Learning Sciences,13:3, 337-386.

Deep, A., Murthy, S., & Bhat, J. (2020). Geneticus Investigatio: a technology-enhanced learning environment for scaffolding complex learning in genetics. Research and Practice in Technology Enhanced Learning. EduMedia sciences.

Herreid, C. F. (2007). *Start with a Story*. Amsterdam University Press.

Jonassen, D. H., & Hernandez-Serrano, J. (2002). Case-based reasoning and instructional design: Using stories to support problem solving. *Educational Technology Research and Development*, *50*(2), 65–77.

Kolodner, J. L. (1992). An introduction to case-based reasoning. Artificial Intelligence Review, 6(1), 3–34.

Kolodner, J. L., Camp, P. J., Crismond, D., Fasse, B., Gray, J., Holbrook, J., Puntambekar, S., & Ryan, M. (2003). Problem-Based Learning Meets Case-Based Reasoning in the Middle-School Science Classroom: Putting Learning by Design(tm) Into Practice. Journal of the Learning Sciences, 12(4), 495–547.

Labster Cytogenetics. (2021). https://www.labster.com/simulations/cytogenetics/

Learn Genetics, Genetic Science Learning Center, University of Utah. (2021). Retrieved on May 31, 2021.https://learn.genetics.utah.edu/content/basics/karyotype/

Riyanto, R., Amin, M., Suwono, H., & Lestari, U. (2020). The New Face of Digital Books in Genetic Learning: A Preliminary Development Study for Students' Critical Thinking. *International Journal of Emerging Technologies in Learning (IJET), 15*(10), 175.

The Biology Project, University of Arizona. (2021). Retrieved on May 31, 2021.http://www.biology.arizona.edu/human_bio/activities/karyotyping/karyotyping2.html

The Science Source SE. (2021). Karyotyping simulation kit. Retrieved on May 31, 2021. https://us.vwr.com/store/product/8878404/karyotyping-simulation-kit

Thermo Fisher Scientific. (2021). Science first human karyotype kit. Retrieved on May 31, 2021. https://www.fishersci.com/shop/products/human-karyotype-kit-karyotyping-kit/s25820a

Training Karyotypes, Sant Joan de Deu hospital, Barcelona. (2021). Retrieved on May 31, 2021. http://trainingkaryotypes.com/.

Virtual cell biology lab, Rutgers University. (2021). Retrieved on May 31, 2021. https://www.ece.rutgers.edu/~marsic/books/SE/projects/ViBE/

Watson, I., & Marir, F. (1994). Case-based reasoning: A review. *The Knowledge Engineering Review*, *9*(4), 327–354.

Design and Deployment of a Mobile Learning Cloud Network to Facilitate Open Educational Resources for Asynchronous Learning

Joselito Christian Paulus M. VILLANUEVA[a*], Mark Anthony V. MELENDRES[a], Catherine Genevieve B. LAGUNZAD[a] & Nathaniel Joseph C. LIBATIQUE[ab]
[a]*School of Science and Engineering, Ateneo De Manila University, Philippines*
[b]*Ateneo Innovation Center*
*joselito.villanueva@obf.ateneo.edu

Abstract: This paper describes the design and deployment of a mobile cloud network that facilitates open educational resource content distribution. The setup utilized clustered single board computers as content, communication and monitoring servers. It was installed in a Public High School where stakeholders, using their mobile devices, were given access to preloaded content via wireless local area network. Initial tests of the mobile cloud showed good network performance. Teachers were randomly selected to evaluate the content validity and delivery of the OER content. Results show that the quality of the network's OER content is very satisfactory. This implementation shows the advantage of mobile cloud computing in the delivery of learning content in remote learning modalities.

Keywords: Mobile learning, open educational resources, remote learning, mobile cloud computing

1. Introduction

The COVID-19 pandemic changed the educational landscape as educators find new and effective ways to reach millions of students worldwide. Educators are forced to employ technologies and strategies to slowly replace traditional face-to-face learning with blended learning methods to narrow the achievement gap (Burgess & Sievertsen, 2020; IAU, 2020; Terada, 2020). Though it had been persisting for some time, this revolution challenges educational institutions in terms of their readiness, technological resources, and human resources.

The digital divide had become more apparent as the pandemic had spread around the world. Equal access to the internet has been a significant issue in the Philippines. Aside from the slowest internet in South East Asia, 45% of Filipinos and 74% of public schools do not have access to the internet. Only about 40% of the country's public high schools have computers with internet connectivity potential. However, only 40% of these schools give students access to and training on using the internet (Jones, 2019). The pandemic challenged the Philippine educational system to evolve to become digital, which most learners are not prepared for (Alipio, 2020).

As a response to this, the Department of Education explored the use of Open Educational Resources (OER). Several countries have already adopted OERs broader use, indicating that this is a future learning solution. Moreover, there is now a growing research interest in maximizing the use of OER's in the current and "post-COVID" education. OERs show promise because of flexibility, cost-effectiveness, and a wide array of applications (Huang, Tlili, Chang, Zhang, Nascimbeni & Burgos, 2020). Since OERs are increasingly adaptable, educators can tweak its substance to fit students' adapting needs and objectives. Instructors can reuse, create, collaborate, and contextualize materials providing wide assortment of materials.

Although OERs are cost-efficient, dissemination of this resource is still a significant challenge. Recent technology developed by the Ateneo Innovation Center (AIC) that focused on Mobile Cloud Computing to cope with high cost, low performing internet access (dela Cruz, Libatique & Tangonan, 2019; Mercado, 2020; Mamaradlo, 2020). Using this technology, students and teachers can access preloaded content in a mesh server that can be accessed through various devices without using the internet. This model has its potential in teaching and learning as it addresses the limited bandwidth and connectivity issues among learners. With preloaded content, the system works as a repository of materials such as videos and documents that can be readily accessed using mobile devices.

This paper describes the design of an offline network to access preloaded open educational content in a mobile cloud network system. The mobile network was deployed in a public high school that aimed to provide offline access to open educational content in the context of asynchronous remote learning modality. This paper also presents the teachers and students perceived quality and mobile learning acceptance.

2. Mobile Cloud Computing

Mobile Cloud Computing (MCC) or mobile cloud technology is an architecture where both data storage and processing take place outside a mobile device. This is based on a new paradigm for the application of mobile where computing occurs in a cloud and that is accessed through wireless connections (Dinh, Lee, Niyato & Wang, 2013). As applied to educational context mobile learning (M-learning) is differentiated from electronic learning (e-learning) and distance learning (D-learning). Rimale, Benlahmar, Tragha & El Guemmat (2016) stressed that students can benefit from M-learning that it facilitates learner interaction, mobile devices are easy to accommodate, mobile devices can be used anytime, more student engagement, mobility and cost effectivity of the devices.

There have been various applications of mobile cloud computing. Most of these applications are internet based or rely on online access for optimization. However, with limited bandwidth to deliver learning content this paper inclines the use of Near Cloud technology. One of the prominent features of the Near Cloud technology is the configuration capability for services such as proxy servers, caching server, database, torrent managers, and communication services which allows optimization of any bandwidth available (dela Cruz, 2018). The nodes of the Near Cloud system are low cost, low power, and low maintenance and can serve as a gateway to the internet. Its architecture has a caching system with terabytes worth of storage which serves as an easily deployable, efficient, and resilient network capable of collecting and transmitting data (Mercado, 2020). A recent study by Talusan, Nakamura, Mizumoto & Yasumoto (2018), demonstrated the implementation of the near cloud architecture for Rural Area Connectivity and Data Processing.

3. Open Educational Resources Adoption

Recently, there is an increase in the literature that referenced Open Educational Practices (OEP) and open resources. Koseglu & Bozkurt (2018), noted that between 2007 to 2017 there is an increase in the trend of peer-reviewed publications that focused on growing awareness of the importance of adopting open practices. This also includes how OERs affect the learning process and teaching practices situated in available courses and platforms with open-source technologies.

Luo, Hosttetler, Freeman & Stefaniak (2020), mentioned that publications about OER include the perception and efficacy of open resources and what hinders the adoption of such. The study stressed that, in terms of quality, OERs are mostly perceived as equivalent to traditional resources and do not harm learning outcomes. Publications, in general, are adept in OER adoption yet studies on barriers in implementation and/or efficacy of these materials are not fully explored.

As practitioners gradually become more accustomed to using OERs, practice about its use is also changing. Adam (2020), stated that as open education is practiced and understood, its implementation is based on the practitioner's "history, worldview, subjectivities, mannerisms, and character" (p.181). This put OER educators in a critical role in the dynamics of the learning environment. Teachers should be viewed as competent in both face-to-face and online instruction and should be literate and skilled users of open resources (Atenas, Haveman & Priego, 2014).

4. Mobile Network Design

The mobile cloud network architecture in this study was designed to be portable, low power and low cost yet capable of the necessary processing capabilities for its implementation. For this, 3 units of Raspberry Pi 3B+ were utilized as servers. It was selected primarily due to its adherence to the network architecture's use case design. It is a low cost, low power and portable Single Board Computer (SBC) and has ample processing power for the network design. Each unit has 1Gb LPDDR2 SDRAM, Gigabit Ethernet port and 2.4GHz and 5GHz 802.11b/g/n/ac Wi-Fi. The Raspberry Pi SBCs were arranged as a cluster to distribute the tasks into three available servers.

The network was designed to host high capacity processing and multiple devices can connect to it at the same time. For this, the Raspberry Pi servers were connected to the network using a D-Link AC2100 Wi-Fi Gigabit router that can provide high bandwidth (up to 300 Mbps for 2.4 GHz and 1733 Mbps for 5GHz). To extend the range covered by the network, Google Wi-Fi Mesh AC1200 (1200 Mbps throughput over 2.4 and 5 GHz) was used. The network design utilized three of these devices for a stable and wireless mesh network.

To facilitate open educational materials and other related resources, each server was installed with open-source software to allow content distribution, communication and network monitoring. The first server was dedicated for content distribution and learning resource management. One of the open resource software installed was Kolibri. Kolibri (2020) is an open-source application that creates an offline server to deliver curated educational resources. It is designed for offline use that packages learning that reduces megabytes worth of data while retaining the original quality. It could be treated as a standalone Learning Management System that features a customizable digital OER curriculum with tools such as exam creation, exercises and differentiated assignment content. It is not necessary for this application to be connected to the internet regularly since updates and latest contents can be synced and shared once it connects with another Kolibri installed device with updated software connected in the same network. Kiwix was also installed in the system. It is an offline content reader for Wikipedia, Project Gutenberg or TED Talks. For this study, an offline Wikipedia was utilized. To facilitate videos and other related media content, the content server was installed with an offline PLEX application. It is a digital media player that facilitates offline access to preloaded music, videos and pictures on a server. Finally, we optimized the server as a collaborative cloud storage with Nextcloud. With this open-source software, the server could facilitate file sharing in an offline local area network.

To facilitate communication among stakeholders within the network, we optimized the second server for real time messaging in case of synchronous learning modality. Rocket.chat was utilized as chat server. This application facilitates collaboration among users thru creating chat rooms for instant messaging and file sharing. The second server was also optimized as a Real-Time Messaging Protocol (RTMP) server. It was incorporated with the network in cases where the server was used as a streaming server and can then be used to stream from multiple sources such as LAN cameras. Since high network traffic is expected in the mobile cloud network, Nginx, as load balancer, uses asynchronous, event-driven web requests. For network monitoring, one of the Raspberry Pi Servers was installed with Nagios Enterprise Monitoring Server (NEMS). NEMS is a pre-configured easy to deploy Nagios Core monitoring software designed to run on microcomputers.

The mobile cloud network was deployed in Kaong National High School in Silang, Cavite, Philippines. As seen in Figure 1, the servers have a wired connection to the gigabit router that can support 2.4Ghz and 5Ghz bandwidth simultaneously. Wi-Fi mesh devices were used to extend the network range. The design utilized three Wi-Fi mesh devices, one of which has wired connection to the router while the rest supported a wireless mesh network. The network is designed to be easily deployable in areas such as school grounds or community centers where students are expected to receive and submit printed learning modules. In these areas, physical distancing and other COVID-19 related protocols are enforced. With the deployment of the network, students are expected to access the materials anytime within a wider range.

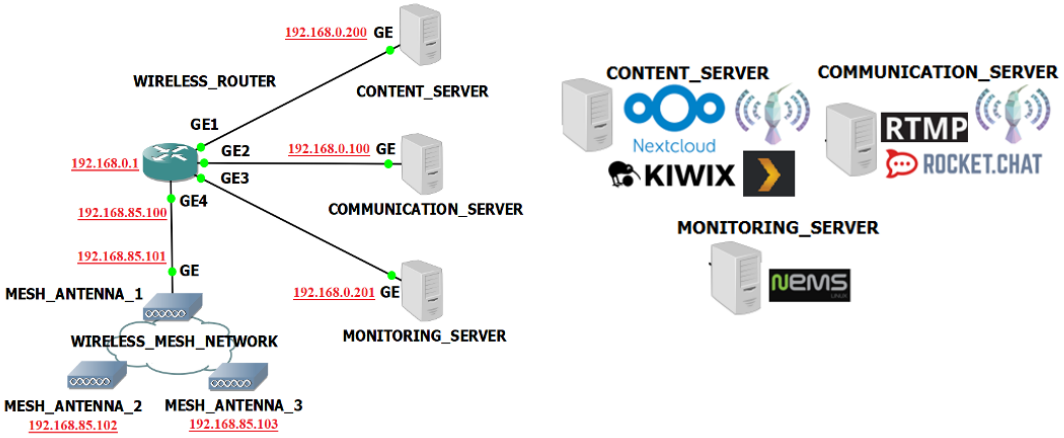


Figure 1. Mobile Network Diagram and Server Design.

5. Characterization of the Mobile Network

5.1 Content Delivery and Management

To facilitate the deployment of the network, teachers and the school administrators were given a demonstration of the network capabilities. Printed manuals were also distributed to the students and parents as well as establishing a social media group for receiving inquiries. It was proposed that during printed module distribution, the students and parents could connect to the network and interact with the supplementary materials. Through the network, they can view and download the supplementary learning resources.

The network was designed to facilitate supplementary materials to the target learners. Following the policy guidelines of the Department of Education, digital file sharing of the modules should only be done by authorized personnel only. Hence, the deployment of the content of mobile cloud network was focused on the distribution and access of supplementary open educational resources. Supplementary materials such as videos, eBooks, simulations and practice assessments were aligned to the Most Essential Learning Competencies (MELC) provided by the Department of Education. Out of the eight subjects in the secondary education curriculum, English, Mathematics, Science and Social Studies supplementary materials were prioritized due to the availability of the resources in the OER repositories.

To be able to connect to the network a client could either connect to the router or the mesh Wi-Fi using any Android, IOS, PC, Linux or Mac devices . The network offers both 2.4 GHz or 5 GHz Wi-Fi bandwidth simultaneously. Once connected, the client can run the applications via opening a browser and typing the IP address of the server that will direct to a landing page of which the user can choose which application to launch in the next tab.

5.2 Network Performance

The initial performance of the network was measured in terms of capacity throughput, network speed and receive signal strength. Four key sampling areas within the school grounds were identified. Figure 2 shows the designated module distribution and waiting areas of which students and parents can connect to the network following the COVID-19 physical distancing protocols. For the first test, capacity throughput was measured for the 2.4 and 5 GHz connection of the mobile cloud network for sixty (60) seconds. Stream tests were done from five (5) to thirty (30) devices. The mobile cloud performs consistently on the 5 GHz compared to the 2.4 GHz band. It was also evident that, in both bands, capacity throughput decreases as more devices are added to the network. However, even at the most number of devices connected that network still showed good throughput performance with 2.52 Mbps on the 2.4 GHz band and 3.15 Mbps on the 5GHz band.

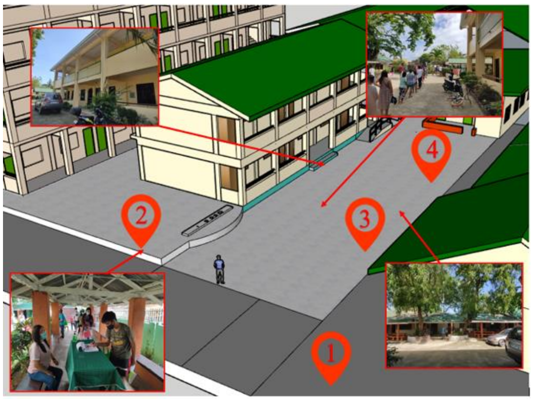


Figure 2. School Site Map, Access Points Location and Module Distribution Area.

The next test measures the bandwidth of the network. Bandwidth was sampled in key areas in the network range and ran for sixty (60) seconds. As seen in Figure 3, comparison of network speeds in the 4 key locations showed high variability. Network speeds vary from 8 Mbps to 43 Mbps with location 3 having the least throughput in the 2.4 GHz and location 2 in the 5 GHz band. Variations in the throughput are possibly caused by adjacent channel interferences of which the current study did not capture. Network bandwidth in the 5Ghz band showed a consistent output. One limitation of the mobile cloud network would be the client's device capacity. Accessing large amounts of data in the network would require a stable and faster network performance that could be limited by the device of the user.

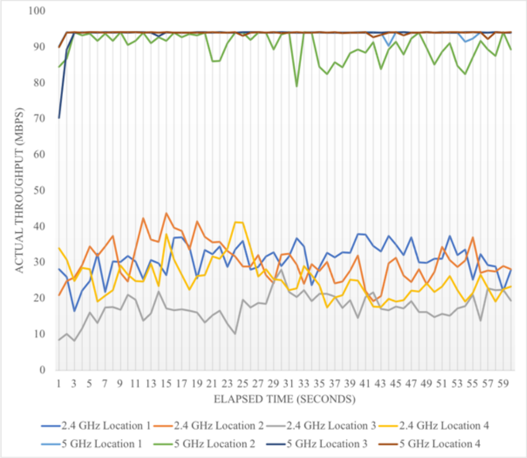


Figure 3. Comparison of Network Speeds of the Four Access Points at 2.4 and 5 GHz.

7. Conclusion

Access to educational resources in times when face-to-face learning modalities are limited puts struggling learners on a disadvantage. The availability of competent learning resources should be made available to all learners despite financial and technological difficulties. This study deployed a platform for the distribution of educational resources that can be locally accessed within the learners' community. The mobile cloud architecture in a mesh network offers a low-cost, low technology requirement, easily deployable and wide range of compatibility systems for the distribution of open educational resources. The system consists mainly of low-cost, low-power consumption and readily available single board computers installed with open source applications for curating and creating learning content. Initial performance tests show that the network architecture is efficient in data transfer and accommodating synchronized usage. It was also evaluated that the network deployed was as more than adequate in the delivery and management of the supplementary learning materials. With the concurrent data provided in this study, future developments on the optimization of the deployed network can enable wider coverage and improved content distribution.

Acknowledgements

The authors would like to acknowledge the Department of Science & Technology – Science Education Institute (DOST-SEI), for their support through the Capacity Building Program in Science and Mathematics Education (CBPSME) and Engineering Research and Development for Technology (ERDT) scholarship grants. The authors would also like to express gratitude to Smart Wireless Engineering Education Program.

References

Adam, T. (2020). Open educational practices of MOOC designers: Embodiment and epistemic location. Distance Education, 41(2), 171–185. https://doi.org/10.1080/01587919.2020.1757405

Alipio, M. (2020). Education during Covid-19 Era: Are Learners in a Less-Economically Developed Country Ready for E-Learning? SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3586311

Atenas, J., Havemann, L., & Priego, E. (2014). Opening teaching landscapes: The importance of quality assurance in the delivery of open educational resources. Open Praxis, 6(1), 29–43. https://doi.org/10.5944/openpraxis.6.1.81

Burgess, S., & Sievertsen, H. H. (2020, April 1). Schools, skills, and learning: The impact of COVID-19 on education. VoxEU.Org. https://voxeu.org/article/impact-covid-19-education

dela Cruz, J. A. (2018). Design of a Disaster Information System using Mobile Cloud Wireless Mesh with Delay Tolerant Network. Ateneo De Manila University.

dela Cruz, J. A., Libatique, N. J., & Tangonan, G. (2019). Design of a Disaster Information System using Mobile Cloud Wireless Mesh with Delay Tolerant Network. 2019 IEEE Global Humanitarian Technology Conference (GHTC), 1–8. https://doi.org/10.1109/GHTC46095.2019.9033450

Dinh, H. T., Lee, C., Niyato, D., & Wang, P. (2013). A survey of mobile cloud computing: Architecture, applications, and approaches: A survey of mobile cloud computing. Wireless Communications and Mobile Computing, 13(18), 1587–1611. https://doi.org/10.1002/wcm.1203

Huang, R., Tlili, A., Chang, T.-W., Zhang, X., Nascimbeni, F., & Burgos, D. (2020). Disrupted classes, undisrupted learning during COVID-19 outbreak in China: Application of open educational practices and resources. Smart Learning Environments, 7(1), 19. https://doi.org/10.1186/s40561-020-00125-8

IAU. (2020). COVID-19: Higher Education challenges and responses—IAU. International Association of Universities. https://www.iau-aiu.net/COVID-19-Higher-Education-challenges-and-responses

Jones, N. (2019). Improving Internet Access In The Philippines (No. 11; CfC Reform Story, p. 9). https://asiafoundation.org/publication/improving-internet-access-in-the-philippines/

Kolibri: A Free, Open Source Education for All | Learning Equality. (n.d.). Retrieved June 16, 2020, from https://learningequality.org/kolibri/

Luo, T., Hostetler, K., Freeman, C., & Stefaniak, J. (2020). The power of open: Benefits, barriers, and strategies for integration of open educational resources. Open Learning: The Journal of Open, Distance and e-Learning, 35(2), 140–158. https://doi.org/10.1080/02680513.2019.1677222

Mamaradlo, J. P. A. (2020). Design And Demonstration Of a Mobile Cloud System For Smart Transportation System Use Case. Ateneo De Manila University.

Mercado, N. A. M. (2020). Design And Demonstration of a Near Cloud System for Digital Education and Disaster Resiliency. Ateneo De Manila University.

Nascimbeni, F., Burgos, D., Campbell, L. M., & Tabacco, A. (2018). Institutional mapping of open educational practices beyond use of Open Educational Resources. Distance Education, 39(4), 511–527. https://doi.org/10.1080/01587919.2018.1520040

Rimale, Z., El Habib, B., Tragha, A., & El Guemmat, K. (2016). Survey on the Use of the Mobile Learning Based on Mobile Cloud Computing. International Journal of Interactive Mobile Technologies (IJIM), 10(3), 35. https://doi.org/10.3991/ijim.v10i3.5672

Talusan, J. P., Nakamura, Y., Mizumoto, T., & Yasumoto, K. (2018). Near Cloud: Low-cost Low-Power Cloud Implementation for Rural Area Connectivity and Data Processing. 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), 622–627. https://doi.org/10.1109/COMPSAC.2018.10307

Terada, Y. (2020). Covid-19's Impact on Students' Academic and Mental Well-Being. Edutopia. https://www.edutopia.org/article/covid-19s-impact-students-academic-and-mental-well-being

EXAIT: A Symbiotic Explanation Learning System

Brendan FLANAGAN[a*], Kyosuke TAKAMI[a], Kensuke TAKII[b], Yiling DAI[a], Rwitajit MAJUMDAR[a] & Hiroaki OGATA[a]
[a]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
[b]*Graduate School of Informatics, Kyoto University, Japan*
*flanagan.brendanjohn.4n@kyoto-u.ac.jp

Abstract: Explainable artificial intelligence has been gaining much attention as systems increasingly make high-stakes recommendation and decisions automatically in a wide range of fields, including education. Meanwhile, research into self-explanation by students as a beneficial intervention to promote metacognitive skills has a long history of research. In this paper, we propose a learning system that aims to bridge understanding between the cyber and physical world by facilitating symbiotic explanation between the EXAIT system and students using the system. A co-evolution cycle of AI recommendation and the self-explanation by students of answer processes is proposed to increase the motivation and awareness of students, and at the same time enhance the effectiveness of the system.

Keywords: Explainable AI, self-explanation, symbiotic learning, recommendation

1. Introduction

In recent years, artificial intelligence has begun to revolutionize many fields, such as: medicine, finance, legal, and education, with decisions being made by models for prediction, estimation, and recommendation. As the use of models in education systems has been increasing, there are greater calls for not only the transparency and interpretation of inner workings of AI systems, but also technologies to explain complex models and the basis of results to end user stakeholders (Wang, Yang, Abdul, Lim, 2019). Much of the research into explainable AI (XAI) has focused on verifying the rationale behind results from such systems, however in education there are also additional benefits of explanations from learning systems that should be considered, such as: students learning from the explanations given by the system. Previous work into intelligent tutoring systems has shown that student's motivation in self-regulated learning with the system can be improved through prompt and feedback mechanisms leading to higher achievement (Duffy & Azevedo, 2015). Therefore, we argue that XAI in education should involve two main facets: explanation to increase student awareness of their course of study, and the explanation of recommendations for the purpose of fostering trust and motivation for continued use of the system.

While the explanation of system recommendations is gaining much attention, research into self-explanation from students in the context of education has a long history since Chi's seminal work on the topic (Chi et al., 1989). In their meta-analysis of 64 research reports on self-explanation and its effect on learning outcomes, Bisra, Liu, Nesbit, Salimi and Winne (2018) highlighted the potential of using self-explanation in a range of learning contexts as a beneficial intervention was suggested, however problems with instructor guided self-explanation was acknowledged. It was recommended that the effect of the intervention could possibly be improved by further investigation into system generated self-explanation scaffolds. Also, the analysis of self-explanations to inform AI driven learning systems, while mentioned to by Chi et al. (1989) is still a topic of ongoing enquiry. There is potential for learning systems to gain insight from detailed descriptions of answer processes that are input as a part of self-explanations by students. Such a system could learn a form of explaining the answer process of a question in Mathematics from high performing students, and then provide the learnt explanations as a scaffold from students that are struggling to perform the same task. Answer process analysis, such as pen stroke input time series analysis in Mathematics to find stuck points (Yoshitake, Flanagan, Ogata, 2020), coupled with self-explanation in the context of the answer process time series could uncover possible weaknesses in prerequisite knowledge for the problem at hand.

In this paper, we propose a system called EXAIT: Educational eXplainable AI Tools, which aims to combine these two aspects of explanation in education into a learning tool that can co-evolve in a symbiotic manner through learner's self-explanation and AI generated explanation. Firstly, the explanation of AI recommendations of possible learning paths from the view point of fostering trust and learner awareness. This then leads to students completing a recommended task, such as: a Mathematics question using a stylus pen to input the working out and final answer into the system for evaluation. The system then prompts the student to self-explain their answer to the task by replaying their answer process interactively, while annotating points in time to indicate what knowledge was applied to overcome sub-problems. Time series analysis is then applied to the self-explanation and answer process data to extract information, such as: backtracking or stuck points where the student had problems during the answer process that could indicate problems with dependent or related knowledge. The final goal of the system is to complete the symbiotic explanation cycle by informing the AI recommendation model from the self-explanation analysis.

2. Related Work

Previous work has proposed theoretical frameworks for symbiotic learning systems (Wu, Yang, Liao, Nian, 2021), where a learner learns from the system while the system also learns from the learner by employing reinforcement learning. However, the main purpose of such platforms and frameworks is to optimize the performance and efficacy of the system by finetuning a model to the behavior of the learners. In the present paper, we propose a symbiotic learning system focusing not only on adjusting the performance of the system to the learner, but also facilitating mutual understanding through the use of explanations by the learner and system to each other.

In the field of learning analytics, the iSTART tutoring system was proposed to tackle problems faced by students when paraphrasing during reading comprehension by supporting the task through the analysis of students' self-explanation in the context of the target text. A combination of NLP methods for self-explanation analysis and scaffolded instructions, and practice of self-explanation were implemented to encourage the development of required skills for the tasks. An automated evaluation algorithm gave students' scores on the quality of their self-explanations, with those that had high relevance of topics to the target text receiving high scores, and those that were off topic or short in length being assigned lower scores. It was found that the system could support self-explanation in different texts of various fields areas (Jackson, Guess, & McNamara, 2010). In the EXAIT system, the analysis of self-explanation will focus less on the automated evaluation of the content, and more on how the system can inform learner models on the strengths and weaknesses. It is intended that the analysis of self-explanation would influence the recommendation of learning paths of the student, thus closing the gap in the explanation co-evolution cycle.

3. EXAIT System

The EXAIT system proposed in this paper builds on the LEAF framework that has been developed to support the distribution of learning materials, collection and automated analysis of learning behavior logs in an open and standards-based approach (Flanagan & Ogata, 2018). The main components of the framework are: Moodle LMS which acts as a hub for accessing various courses; the BookRoll reading system for learning material and quiz exercise distribution; an LRS for collecting learning behavior logs from all of the components; and the LAView learning analytics dashboard to provide feedback to students, teachers and school administrators. This framework enables the real-time collection, analysis of learning behavior and feedback direct to stakeholders. An overview of the EXAIT concept is shown in Figure 1, with ⑤ LEAF platform being the foundation of the whole system. On top of this there is an abstracted model layer made up of ③ data-driven and ④ model-driven aspects. The model-driven aspect comprises of two main models: the knowledge model that represents the structure in the form of a knowledge graph of concepts that are learnt within a course, and the student model to monitor learning progress of individuals by analyzing data collected in the LEAF platform. The data-driven aspect comprises of an evidence model that guides the system based on the past performance of interventions. The interaction layer at the top of the diagram shows the symbiotic nature of the system while students are learning. There are two key aspects of the interaction: the system's ability to ① explain recommendations or decisions that have been made to the stakeholders, and reciprocal explanation by the students in the form of ② self-explanation of answers that they have submitted while using EXAIT. This reciprocal explanation between the system and students forms the basis of the symbiotic process that is the fundamental core of EXAIT.

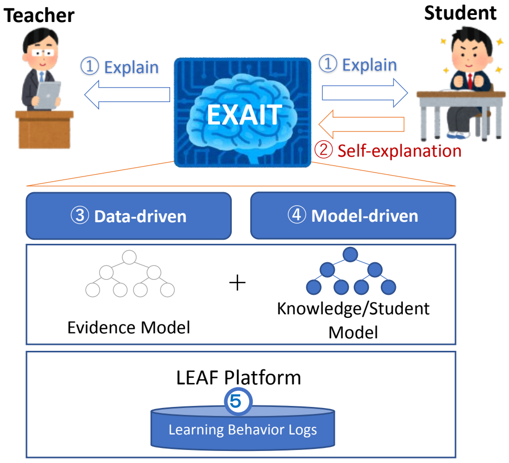


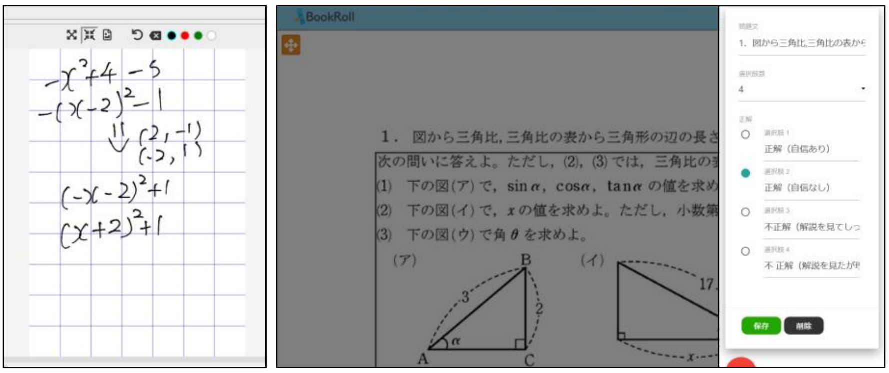
Figure 1. A High-level Overview of the EXAIT System.

The evaluation of the proposed system currently focuses on English and Mathematics education in Japanese secondary schools. The reason for this is that it is currently the target of the Japanese governments GIGA-school program to provide one computer per student to all children in compulsory education and mainly targeting secondary schools initially. Brod (2020) also highlights evidence from previous research showing that there is a favorable effectiveness of self-explanation when used in secondary school or higher. This confirms that the selection of secondary schools is a good fit with the proposed methods in the EXAIT system. In this paper, we will present the implementation of the EXAIT system for Mathematics education.

3.1 Recommendation Explanation in Mathematics

Exercise books used in the Mathematics course were uploaded to BookRoll and multiple-choice quiz questions were created to enable the collection of answers in learning log data. Students are also required to submit the working out that lead to the answer of the question as a hand written memo in the BookRoll interface as shown in Figure 2. The recommendation of learning paths in Mathematics has been popularized by the application of Bayesian Knowledge Tracing (BKT) in intelligent tutoring systems (Fischer et al., 2020), and is employed to model the degree of mastery of each mathematical skill in the EXAIT system based on the analysis of answers in learning log data. Question recommendations are made based on the probability that the student will correctly answer a question as determined by the BKT model, with extremely high or low probability questions having less weight in the recommendation.

We propose two main aspects from which recommendations should be explanation: Learner-centric and content-centric. The personal experiences of a learner can be extracted from learning logs to generate learner-centric explanations based on their past activity. An example of this could be that the learner has answered correctly questions on a skill that are prerequisites of the current recommended question. The EXAIT system collects evidence of previous cases of effective recommendation in the evidence model, which could also be used to explain to the learner about possible future learning path trajectories after completing the recommended question. Content-centric explanation can be generated by interpreting the parameters of the BKT model that has been trained, such as: the probability of transitioning from a state of unknown to mastery, the degree of forgetfulness or change that the answer might be a careless mistake. The clustering of these parameters will be analyzed to identify different types of questions for explanation, such as: the basic recall of knowledge in fundamental questions, to the application of knowledge in advanced questions. It is anticipated that this explanation will increase the learner's awareness of not only the type of question that is being recommended, but also the current phase of study.



1． 図から三角比, 三角比の表から三角形の辺の長さ

次の問いに答えよ。ただし，(2)，(3) では，三角比の表

(1) 下の図(ア)で，sin α，cosα，tanα の値を求め

(2) 下の図(イ)で，x の値を求めよ。ただし，小数第

(3) 下の図(ウ)で角 θ を求めよ。

問題文：1．図から三角比,三角比の表から

選択肢数：4

正解

○ 選択肢 1：正解（自信あり）

● 選択肢 2：正解（自信なし）

○ 選択肢 3：不正解（解説を見てしっ

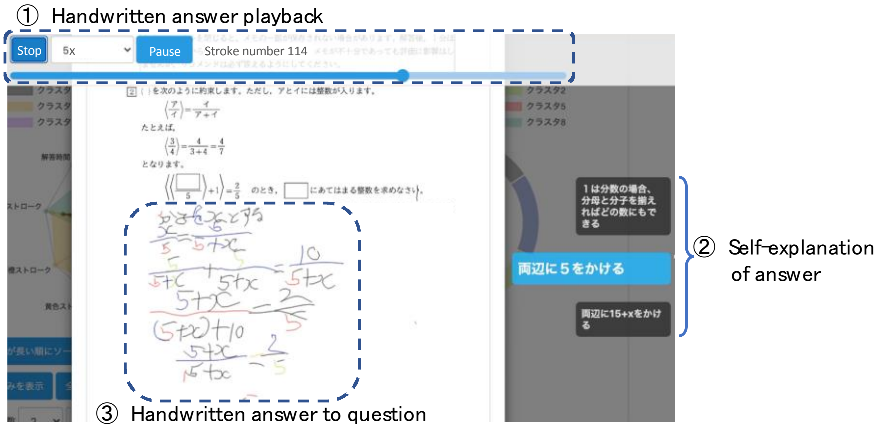
○ 選択肢 4：不 正解（解説を見たが理

Figure 2. Handwritten Answer as a Memo (Left), Playback, and Self-explanation Input.

3.2 Answer Self-Explanation in Mathematics

As a part of the system, students are asked to explain their handwritten answers to questions that have been recommended by the system. The current user interface as shown in Figure 3 also includes the time delay analysis of pen strokes in the handwritten answer to indicate where students paused during the answer process (Yoshitake, Flanagan, Ogata, 2020). The student can playback and reflect on their answer by using the ① handwritten answer playback interface at the top. The rate of playback can be selected and users can also use the jump bar to skip to different parts of the answer process. When the playback is paused, the student has the option to create a new explanation for the particular step in the answer process. During playback the ② self-explanations of answers are displayed on the right of the screen and scroll through each of the explanations highlighting the current one as seen in Figure 3.

Previous research has shown that self-explanations can promote meta-cognitive skill use, as the students diagnose their answers, drawing on previously learnt knowledge that was applied to solve sub-problems (Chiu & Chi, 2014). It can also increase a student's awareness of learning by generating and completing self-explanation, using examples to convey new explanations of knowledge that has been acquired. The outcome of using examples to create self-explanations is that the student constructs more sophisticated knowledge through the process. However, while working with the class teachers, it became apparent that most students do not regard this task as worthwhile and it is viewed as more of a hassle. Also, students that have not grasped the required knowledge struggle with producing explanations when compared to high achieving students.


① Handwritten answer playback
Stop
5x
Pause
② Self-explanation of answer
③ Handwritten answer to question
クラスタ2
クラスタ5
クラスタ8

Stroke number 114

2 ⟨ ⟩を次のように約束します。ただし，アとイには整数が入ります。

$\left\langle \frac{ア}{イ} \right\rangle = \frac{イ}{ア+イ}$

たとえば，

$\left\langle \frac{3}{4} \right\rangle = \frac{4}{3+4} = \frac{4}{7}$

となります。

$\left\langle \left\langle \frac{\square}{5} \right\rangle + 1 \right\rangle = \frac{2}{5}$ のとき，□ にあてはまる整数を求めなさい。

1は分数の場合、分母と分子を揃えればどの数にもできる

両辺に５をかける

両辺に15+xをかける

Figure 3. Handwritten Answer Analysis, Playback, and Self-Explanation Input.

Students that have not acquired sufficient knowledge to create self-explanations for their answer to a question may struggle or generate poor examples. Therefore, it may be necessary to support a student in the task, and could involve restricting the method of self-explanation as suggested by Wylie & Chi, (2014) ranging from open to limited respectively: Open-ended self-explanation (natural language); Focused self-explanation; Scaffolded self-explanation; Glossary/Resource-based self-explanation; Guided self-explanation. Currently, the EXAIT system has implemented an open-ended self-explanation interface with a high degree of freedom of expression. The self-explanations are input at points in time where the student deems appropriate without any prompting or guidance from the system. While this might be effective in a classroom where the instructor can guide the students through the process, there are potential problems with depending on such support as highlighted in previous research by Bisra, Liu, Nesbit, Salimi & Winne (2018). In the next iteration of the system, we plan to implement a guided process map based self-explanation interface that will provide students with predetermined keywords that are generated by the system. These keywords will be selected based on the required knowledge to complete the task. The student will then be prompted to arrange the keywords in order of the answer process and assign appropriate points in time where the knowledge was utilized.

4. Preliminary Study

In this section we describe two preliminary studies that were done to assess the use of recommendations provided the system without explanation. A recommender for both English and Mathematics was deployed to a secondary school in Japan. The students that participated in the study have past experience of using the LEAF system for at least the past academic year. The English reading recommender by Takii, Flanagan & Ogata (2021) was deployed as a book recommendation system for extensive reading in ESL classes. It is a knowledge map recommender method based on the vocabulary knowledge structure proposed by Flanagan et al. (2019). It was made available through two different dashboard systems: LAView which is a part of LEAF and GOAL a dashboard for Self Direction Skills (Majumdar et al., 2018). The Mathematics recommender proposed in previous sections was deployed only in LAView as GOAL had not previously been introduced to the target classes. Students were provided an orientation on how and why they should use the recommenders by the teacher of the course. As shown in Figure 4, initially there was high usage of the recommenders regardless of the subject or the system that was used for access. However, as time passed the usage of the recommender reduced even though the students were still engaged in studies. It should be noted that the time frame of the two studies is different in duration, with the English recommender being conducted over two weeks, and Mathematics over two months.

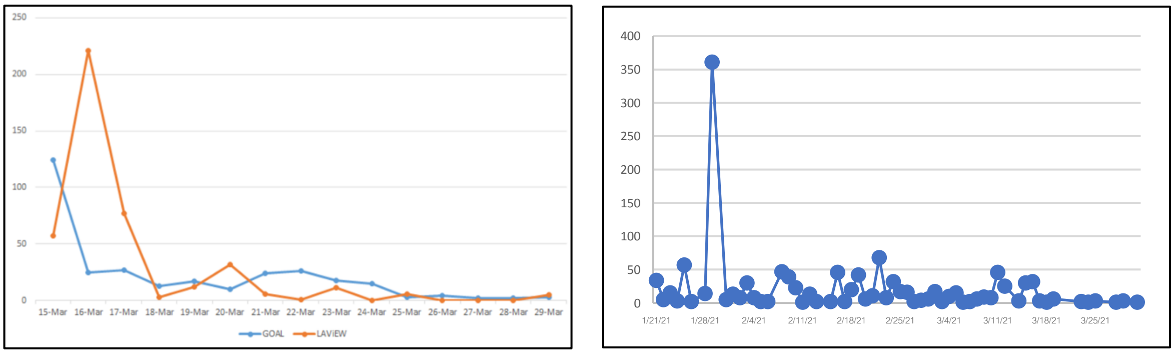


Figure 4. Recommender Usage (access frequency): extensive reading book recommendation for English class (left), question recommendation for Mathematics class (right).

5. Conclusion and Future Work

Recently, XAI has been gaining much attention in many different fields. In the context of education, we propose that there are some aspects of XAI that are unique to the field, such as the possibility of students learning from the explanations of a recommender. The implementation of AI system explanation and student self-explanation provide an opportunity to create a symbiotic learning system, which we proposed called EXAIT. In the present paper, we outline the current recommendation explanation and self-explanation components of EXAIT, and discuss possible future directions and challenges that remain. Two preliminary studies were conducted to see the use of recommendations in English and Mathematics courses without explaining the decisions to the students, and it was found that usage tapered off over time. In future work, we plan to formally evaluate learner-centric and content-centric methods of recommendation explanation. We also plan to implement and evaluate the effectiveness of computer-generated scaffolding for self-explanations by students.

Acknowledgements

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) 20H01722, JSPS Grant-in-Aid for Scientific Research (Exploratory) 21K19824, JSPS Grant-in-Aid for Scientific Research (S) 16H06304 and NEDO JPNP20006 and JPNP18013.

References

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing Self-Explanation: a Meta-Analysis. *Educational Psychology Review*, *30*(3), 703-725.

Brod, G. (2020). Generative Learning: Which Strategies for What Age?. *Educational Psychology Review*, 1-24.

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science*, *13*(2), 145-182.

Chi, M. T. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. *Advances in instructional psychology*, *5*, 161-238.

Chiu, J. L., & Chi, M. T. (2014). Supporting self-explanation in the classroom. *Applying science of learning in education: Infusing psychological science into the curriculum*, 91-103.

Duffy, M. C., & Azevedo, R. (2015). Motivation matters: Interactions between achievement goals and agent scaffolding for self-regulated learning within an intelligent tutoring system. *Computers in Human Behavior*, *52*, 338-348.

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education*, *44*(1), 130-160.

Flanagan, B., Ogata, H. (2018). Learning analytics platform in higher education in Japan, *Knowledge Management & E-Learning: An International Journal*, *10(4)*, 469-484.

Flanagan, B., Chen, M. A., Lecailliez, L., Majumdar, R., Akçapinar, G., Ocheja, P., & Ogata, H. (2019). Automatic vocabulary study map generation by semantic context and learning material analysis. In *Proceedings of the 27th International Conference on Computers in Education (ICCE2019)* (pp. 698-702).

Jackson, G. T., Guess, R. H., & McNamara, D. S. (2010). Assessing cognitively complex strategy use in an un-trained domain. *Topics in Cognitive Science, 2*, 127–137.

Majumdar R., Yang, Y.Y., Li, H., Akçapınar G., Flanagan B., Ogata H. (2018). GOAL: A System to Support Learner’s Acquisition of Self Direction Skills, In *Proceedings of the 26th International Conference on Computers in Education (ICCE2018)*(pp. 406-415).

McNamara, D. S., Boonthum, C., Levinstein, I. B., & Millis, K. (2007). Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 227–241).

Takii, K., Flanagan, B., & Ogata, H. (2021). An English Picture-book Recommender System for Extensive Reading Using Vocabulary Knowledge Map. *Companion Proceedings of the 11th International Conference on Learning Analytics and Knowledge*.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019, May). Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1-15).

Wu, J. Y., Yang, C. C., Liao, C. H., & Nian, M. W. (2021). Analytics 2.0 for Precision Education: An Integrative Theoretical Framework of the Human and Machine Symbiotic Learning. *Educational Technology & Society*, *24*(1), 267-279.

Wylie, R., & Chi, M. T. H. (2014). The Self-explanation principle in multimedia learning. In R. E. Mayer (Ed.), The Cambridge handbook of multimedia learning (2nd ed., pp. 413-432).

Yoshitake, D., Flanagan, B., & Ogata, H. (2020). Supporting Group Learning Using Pen Stroke Data Analytics. In *Proceedings of the 28th International Conference on Computers in Education Conference Proceedings (ICCE2020)* (pp. 634-639).

TinkerBot: A Semi-Automated Agent for a Learning Environment Based on Tinkering

Shruti JAIN*, Ashutosh RAINA & Sridhar IYER
IDP in Educational Technology, Indian Institute of Technology Bombay, India
*jainshruti429@gmail.com

Abstract: Tinkering is an effective approach to develop Engineering Design Skills. Tinkering based robotics kits can aid in nurturing tinkering and associated design skills. Due to a limited understanding and experience, novice designers face several challenges like access to basic information via rudimentary sources like manuals leading to complications in accessing basic information. Design fixation is another challenge that limits the learners in utilizing such robotics kits to their full potential. Though a mentor can address these challenges by scaffolding and aiding reflection, a mentor supported by a semi-automated agent can have a tremendous impact on a seamless familiarization and with the robotics kits. In this paper, we present an idea of a chatbot based Semi-Automated agent as a companion for tinkering kits. After analyzing interactions between a mentor and a participant in the robotics workshops we have classified the interactions and used them to develop a scaffolding logic to automate certain routines of interactions using the chatbot. Such a chatbot can act as a scaffolding agent as well as a companion for journaling and also open the possibilities to remote or virtual mentoring.

Keywords: Tinkering, Chatbots, Engineering Problem Solving, Reflection Journal

1. Introduction

It is increasingly important that the next generation of students must acquire problem-solving and critical thinking skills to succeed in the 21st century (Afari & Khine, 2017). Tinkering is a productive approach to learn problem-solving skills. It can be defined as the process of trying an idea, evaluating and refining the solution. While solving problems with the LEGO robotics kit, students are involved in the tinkering process (Resnick & Robinson, 2017) and such kits are known to help learners in developing innovative ideas, disruptive thinking and higher-order learning skills (Afari & Khine, 2017).

Literature on nurturing tinkering with robotic companions (Raina et al., 2019); and on reflection (Patel & Dasgupta, 2019) suggest that scaffolding in the form of reflection from a mentor plays an important role in learning to solve engineering design problems and nurturing tinkering behavior. Nowadays, Chatbots are gaining popularity in a wide range of applications especially in systems that provide intelligent support to the user (Colace et al., 2018). They are considered as cheap and easy to use educational tools, which are closer to nowadays students and adequate to modern styles of learning (Georgescu, 2018). There is very little research on using Chatbots in the context of problem-solving or engineering design.

While working with Robotics Kits, novice designers face several challenges. Firstly, they lack experiential knowledge of working with the available materials or the means of gathering such knowledge as and when required. This lack inhibits their ability to tinker. This could be addressed, by introducing a mentor companion who gives *Just In Time Information* and *Just In Time Tinkering Triggers (JITI and JIT3)* (Raina et al., 2019). The second problem they face is fixation. Design fixation, which is defined as adherence to one's own design, is a common problem in engineering. Engineers tend to stick to the same idea throughout the design cycle (Kershaw et al., 2011). Through various studies, it was found that reflection can help overcome fixation. Various elements including externally prompted questions can act as *Triggers of Reflection* (Patel & Dasgupta, 2019). Reflection may also be actively facilitated through interactive journal writing (Daive, 1997).

Previous solutions like *TinkMate* (Raina et al., 2019), to build an automated tinkering companion, solved only the first problem. Through the robot, COZMO, the authors could give conversational triggers through audio and behavioral triggers through animated facial expressions, but it was limited to the triggers which were coded into it, there was no room for creating design journals or logging participant activity, it could not be used beyond a workshop setting and all the interactions with the mentor remained outside the app. These limitations can be addressed through chatbots. Chatbots can rely on programmed logic or use Artificial Intelligence (AI) and Natural Language Processing (NLP) to interpret user interactions providing quick and fast responses, seamlessly. (Colace et al., 2018). In this paper, we attempt to develop a semi-automated scaffolding agent in form of a chatbot, that can be used with engineering design kits like LEGO Mindstorms in workshop settings and help novice tinkerers learn problem-solving skills in the context of tinkering.

2. TinkerBot

To aid reflection and overcome challenges of getting stuck while working with Robotics Kits we designed a chatbot, “TinkerBot”. TinkerBot is based on Slack APIs and is primarily designed to assist mentor participant interactions in a workshop setting. Mentors can send prompt, trigger and check on participant progress in a partially-automated manner in a conversational mode. TinkerBot provides the participants with an interactive logging journal and allows them to see the problem statement, required resources; and interact with mentor on the same platform. It maintains participant states by monitoring their activity, this can be leveraged to offload mentor checks, allowing the mentor to be engaged with multiple participants at the same time. So, in theory, using TinkerBot, mentor can keep a bird-eye view of the participants and get involved with them in a seamless manner. TinkerBot has three major components the Slack API infrastructure, Data Store and Scaffolding Logic.

Slack API infrastructure: Slack offers a wide range of APIs for designing powerful Chatbots. In addition to automated conversations, they can send scheduled messages and share files and documents. They are also equipped with interactive messages that can be leveraged for structured conversations. They involve forms that can be used for collecting user feedback or creating a log journal. They can store and retrieve data from databases and help in monitoring user activity.

Data Store: The data store has an organized list of challenges divided into phases based on the workshop sequence or complexity. It also has resources in the form of files, images and weblinks. These resources are mapped with tasks so that only the resources relevant for a given task are provided to the participant. Individual participant data: their current phase, challenge state, collectively referred to as participant state and logging journal are also stored in the data store.

Scaffolding Logic: The scaffolding logic governs the automated and semiautomated prompts and triggers, based on progression of the participant in a given challenge, time-based events, participant’s activity on the app or commands explicitly sent by the mentor. Certain routines of

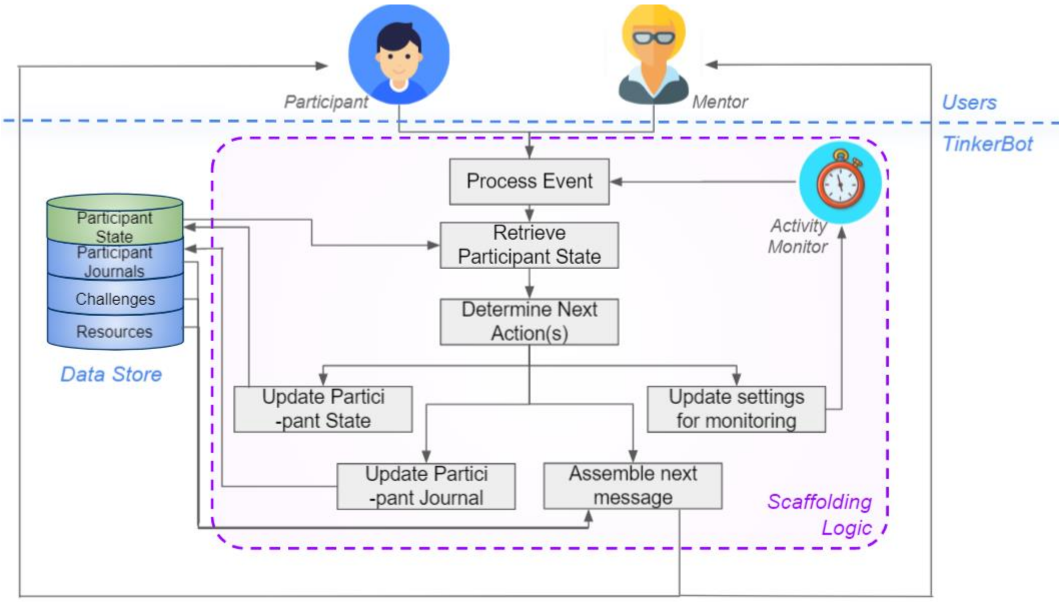


Figure 1. Scaffolding Logic and its Interaction with Users, Data Store and Monitoring App.

mentor-participant interactions identified based on the classification, presented in the next sections, were automated in form of conversations. The Scaffolding Logic is designed as a decision tree. When an event is raised, in form of a message from the users, or internally by the activity monitor, the logic takes this event and participant state and as input, computes the response as seen in figure 1 and makes updates if any. The required resources or information is then pulled from the data store to structure the output which could be the next message sent to either the mentor or the participant.

In the current state of development, we have coded several routines into TinkerBot, a few of them are shown in figure 2. TinkerBot can manage the complete flow of a challenge from introducing the problem statement to completion. When a new participant is registered, TinkerBot sends an interactive message as shown in figure 2(a), it introduces itself as a friendly companion and then talks about different components of the LEGO Mindstorm Kit. Figure 2(b) shows the routine after the generic introduction: TinkerBot waits for the participant to hit "Ready" after they have gone through the shared resources and then the bot would send the problem statement along with few detailed resources required to solve that problem. A few intuitive commands are added for both participant and mentor. Figure 2(b) shows the commands that can be used by a participant, for e. g., "help" displays the list of commands and "ask-mentor" sends a notification to the mentor so that they can join and help the participant. Figure 2(c) shows *a* form to add an entry to the participant's journal, participant can also add pictures to it by attaching files in the chat. Figure 2(d) shows another form, *using which* the mentor can select the next task(challenge) for the participant, they can select from the list or create a custom challenge.

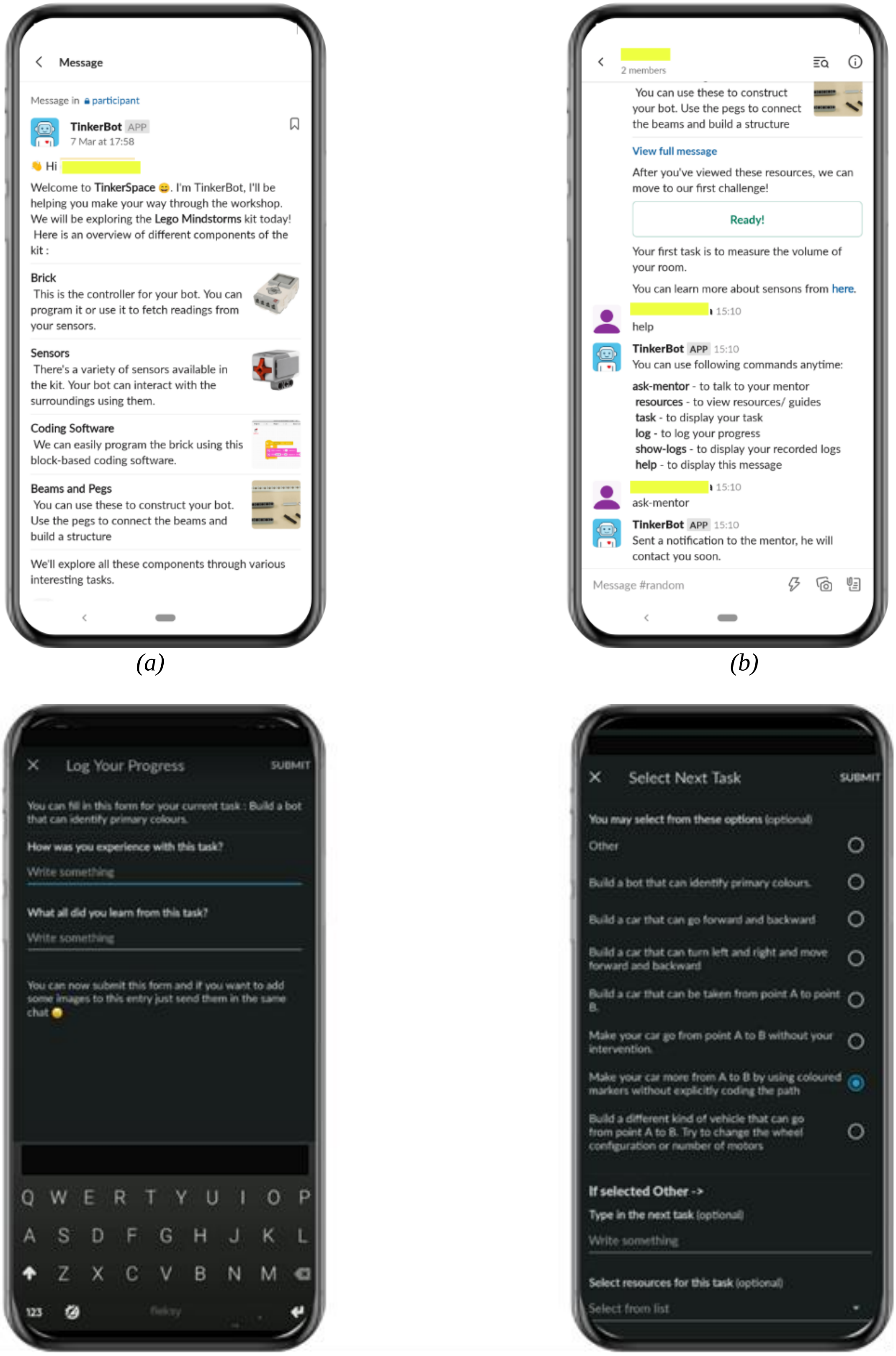
3. Exploration of Participant Interactions

To understand the participant's interaction with mentors and other resources, video data from a tinkering workshop was analyzed. The workshop was conducted as a trial study for designing a learning environment for nurturing tinkering using the Lego Mindstorms EV3 kit. A first-year engineering undergraduate having little experience in robotics volunteered for the study. The pedagogy required the study to be divided into 3 phases- "Explore", "Solve" and "Evolve" (Raina et. al., 2018). Each phase was 3 hours long and comprised of multiple challenges to be solved by the participant, the complexity of challenges increased in every phase. The initial challenges were to measure the volume of the room, identify colors with sensors to building a four-wheel bot to be controlled by remote and eventually making the bot autonomous. The participant was free to search for resources on the internet and a mentor (researcher and experienced tinkerer) was present to guide the participant.

The mentor was to interact only on the initial two days, hence video/audio data of 3 hours for two days of participant activities and interaction with the mentor was observed. The study was set in the research lab where the participants were in the study room and observers sat in the observation room to observe both participant and mentor behavior from a one-way mirror. Participant interactions with the system were triangulated with the screen recording. An informal and open-ended interview of the participant was conducted at the end of each phase. The observations were followed by a round of discussions between the observer and mentor to confirm the intent and potential classification of data. Different stages of the participant's problem-solving process were observed. The instances of mentor-participant interaction and events initiating them were identified and further classified based on intent into different types of prompts/triggers given by the mentor. Finally, routines were derived from frequently occurring sets of conversations which could be automated using a Chatbot.

4. Findings

Through five challenges, spread out across three phases of the study, we observed patterns followed by the participant while solving the challenges. After the challenge was introduced, the participant's trajectory started with finding information about components of the kit, followed by drafting an initial design. With the design ready, the participant started constructing/programming as required. This was followed by testing and closure. These stages may not occur in sequence, for e. g., the participant may revisit resources while constructing the required structure. Most of these stages were more explicit in the case of a novice designer, as seen in “Explore” and early part of the “Solve” phase, but as they gain experience, the process of reflection and building become intertwined.



(a)

Message

Message in 🔒participant

TinkerBot APP
7 Mar at 17:58

👋 Hi [redacted]

Welcome to **TinkerSpace** 😄. I'm TinkerBot, I'll be helping you make your way through the workshop. We will be exploring the **Lego Mindstorms** kit today!
Here is an overview of different components of the kit :

Brick
This is the controller for your bot. You can program it or use it to fetch readings from your sensors.

Sensors
There's a variety of sensors available in the kit. Your bot can interact with the surroundings using them.

Coding Software
We can easily program the brick using this block-based coding software.

Beams and Pegs
You can use these to construct your bot. Use the pegs to connect the beams and build a structure

We'll explore all these components through various interesting tasks.

(b)

[redacted]
2 members

You can use these to construct your bot. Use the pegs to connect the beams and build a structure

View full message

After you've viewed these resources, we can move to our first challenge!

Ready!

Your first task is to measure the volume of your room.

You can learn more about sensons from here.

[redacted] 15:10
help

TinkerBot APP 15:10
You can use following commands anytime:

ask-mentor - to talk to your mentor
resources - to view resources/ guides
task - to display your task
log - to log your progress
show-logs - to display your recorded logs
help - to display this message

[redacted] 15:10
ask-mentor

TinkerBot APP 15:10
Sent a notification to the mentor, he will contact you soon.

Message #random

(c)

✕ Log Your Progress SUBMIT

You can fill in this form for your current task : Build a bot that can identify primary colours.

How was you experience with this task?
Write something

What all did you learn from this task?
Write something

You can now submit this form and if you want to add some images to this entry just send them in the same chat 😊

(d)

✕ Select Next Task SUBMIT

You may select from these options (optional)

- Other ○
- Build a bot that can identify primary colours. ○
- Build a car that can go forward and backward ○
- Build a car that can turn left and right and move forward and backward ○
- Build a car that can be taken from point A to point B. ○
- Make your car go from point A to B without your intervention. ○
- Make your car more from A to B by using coloured markers without explicitly coding the path ●
- Build a different kind of vehicle that can go from point A to B. Try to change the wheel configuration or number of motors ○

If selected Other ->
Type in the next task (optional)
Write something

Select resources for this task (optional)
Select from list

Figure 2. (a) Interactive Introductory Message received by the Participant. (b) Participant can discuss with mentor on the same platform. (c) Form to create a logging journal. (d) Mentor can select next challenge from the list or can create a custom challenge.

4.1 Events Initiating Mentor-Participant Interaction

Mentor-participant interaction was defined as a continuous set of dialogue that occurred between them. Multiple instances of such interactions were observed and classified based on the initiating events.

Participant asks a question: After the problem is introduced, if the participant is completely unaware of components/functionality they directly seek the mentor's help. Example: *P: I have to use this(brick)... I need to see how to get started… I don't know…*

Participant is Fixated: It is observed at several stages, while browsing resources, constructing/programming or testing, that the participant is unable to proceed further or is fixated(stuck), the mentor helps by posing reflective questions or giving hints to the participant.

- *Stuck while browsing resources:* Novice learners often get lost among the wide range of videos and manuals available on the internet. In literature (Raina et. al., 2019) this is referred to as a switch in context, and it takes up a lot of valuable time and causes frustration. *Example: M: Are you looking for something in particular? P: I am not sure how to ...*
- *Stuck while constructing something:* As seen in literature (Patel & Dasgupta, 2019) and our data, a pre-built or demo structure serves as a good reference for the participant, but given the range of building blocks and unfamiliarity with the kit, the participant may not be able to find correct parts for their design. Mentor helps the participant to build the structure from available components. Example: *M: Do you really need it (building component) for your design?*
- *Stuck on programming software:* A participant with limited exposure to computational thinking (use of conditional statements or loops) often struggles while trying to program the Lego EV3 brick. The mentor provides alternates and debugging tips. Example: *M: Why do you think the loop is stuck? P: I am not sure. M: Try tracking the value of this variable.*
- *Stuck on a Design:* Even after realizing that their initial design is not adequate, the participant continues to adhere to it, also referred to as Design Fixation. At this point, the participant needs to reflect on their actions that lead to this current state, rectify the step where the solution went wrong or think of a new approach altogether. To trigger such reflection, the mentor poses a reflection question. Example: *M: Why do you think the bot is not turning? What is making it behave this way?*

Time-based events: Time also acts as a trigger for reflection (Patel & Dasgupta, 2019). When given a time prompt, participant was observed realizing being stuck working at the same step/design/idea for a long time. This realization was followed by reflection on their decisions to figure out a faster approach.

4.2 Prompt Types

On further analysis of mentor-participant interactions, a classification was made among different types of prompts given by the mentor.

Scaffold for Initiating a plan: Novice participants may need some scaffolding to arrive at an initial plan. These can be detailed discussions in which the mentor helps the participant to relate to their previous experiences or real-life examples which can be used to start with an initial plan for the given challenge. Example: *M: In your previous project you had 2/4 motors?*

Triggers to tinker: Such prompts were required during the ideation phase if participant got obsessed with the details of the plan or the working of sensors or bricks. In such cases, the mentor prompted the participant to start building something or to start tinkering. Example: *M: Why don't you try building the bot like you said initially?*

Direct Information: When asked by the participant or while introducing a new challenge, the mentor gives out direct information. Example: *M: These are some sensors you can use…*

Active Mentor Intervention: These were instances when the mentor was actively helping the participant. They are usually ill-structured and involve large sequences of dialogues, for instance, mentor demonstrating some brick functionality or construction of a small structure.

Triggers for reflection: Both time based and contextual prompts act as triggers for reflection (Patel & Dasgupta, 2019). Mentor was seen giving these prompts at various stages of a challenge, these helped the participant to make better design decisions. and traceback their steps and identify alternative approaches. Example: *M: Based on the height, do you think the length is correct?*

5. Study Plan

To understand the impact of a bot-based mentor for a workshop setup we initially aim at understanding the interaction of the automated system such as TinkerBot as an aid in the tinkering process by scaffolding reflection. We plan to do this by answering questions like, (i) How does interactive logging and articulation help in triggering reflection? (ii) How do prompts given by TinkerBot, help to solve engineering design problems? (iii) How does interaction vary as the participant gains expertise? Similar to the exploratory study, this study will be conducted with a novice tinkerer but the mentor will not be physically available and shall only interact via TinkerBot. Video data of the participant's activities and will be captured along with the app interaction log data. The study will be followed by interviews with the mentor and the participant to understand their experience and the challenges they faced while using TinkerBot. This study will help us understand the useability of such a solution in a workshop setting. Due to the closure of lab facilities because of COVID-19, the initial plan of lab-based studies was halted and now we plan to conduct the studies in a remote workshop mode where the participants will work remotely and the mentor will communicate via TinkerBot.

6. Conclusion

In this paper, we explored the possibility of developing an automated scaffolding agent to nurture tinkering behavior in novice designers. We attempted to understand the process of solving a challenge and did an analysis of mentor-participant interaction, we identified events that initiate a conversation between mentor and participant and classified different types of prompts given by the mentor. These states, events and prompts together helped us to develop routines of conversations that were coded into the chatbot in form of decision trees. We plan to conduct studies in future, using the proposed chatbot as the scaffolding agent and explore further possibilities. While a mentor is irreplaceable but developing a hybrid model can prove to be very efficient as a mentor's presence is limited. Chatbot's conversational nature can allow it to act as a companion which is limited with a mentor. Through TinkerBot, a single mentor can manage multiple participants, especially helping the mentor offload various tasks.

References

Afari, E., & Khine, M. S. (2017). Robotics as an educational tool: Impact of lego mindstorms. International Journal of Information and Education Technology, 7(6), 437-442.

Resnick, M., & Robinson, K. (2017). Lifelong kindergarten: Cultivating creativity through projects, passion, peers, and play. MIT press.

Raina, A., Murthy, S., & Iyer, S. (2019, December). Designing TinkMate: A Seamless Tinkering Companion for Engineering Design Kits. In *2019 IEEE Tenth International Conference on Technology for Education (T4E)* (pp. 9-14). IEEE.

Patel, A., & Dasgupta, C. (2019, July). Scaffolding structured reflective practices in engineering design problem solving. In 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT) (Vol. 2161, pp. 287-289). IEEE.

Colace, F., De Santo, M., Lombardi, M., Pascale, F., Pietrosanto, A., & Lemma, S. (2018). Chatbot for e-learning: A case of study. *International Journal of Mechanical Engineering and Robotics Research*, *7*(5), 528-533.

Georgescu, A. A. (2018). Chatbots for education–trends, benefits and challenges. In *Conference proceedings of» eLearning and Software for Education «(eLSE)* (Vol. 2, No. 14, pp. 195-200). " Carol I" National Defence University Publishing House.

Kershaw, T., Holtta-Otto, K., & Lee, Y. S. (2011). The effect of prototyping and critical feedback on fixation in engineering design. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, No. 33).

Daive, L. (1997). Facilitating reflection through interactive journal writing in an online graduate course: A qualitative study. *Journal of distance education*, *1*(12), 103-126.

Raina, A., Murthy, S., & Iyer, S. (2018, July). " Help Me Build": Making as an Enabler for Problem Solving in Engineering Design. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)* (pp. 455-457). IEEE.

A Mixed Study to Understand Taiwanese Children's Preference for A Mobile Game

Yi Chen WANG, Wei Tung Nien & Joni Tzuchen Tang*
Graduate Institute of Applied Science and Technology, National Taiwan University of Science and Technology, Taiwan (R.O.C.)
*jttang0@mail.ntust.edu.tw

Abstract: This mixed study focused on an educational mobile game, "Lily's Daily Life," to explore children's preference by big data in 8216 children and behavior observation for 12 children. The data result shows the game most attracted 3-5 years old children. The highest frequency of play is 3 years old. The results presented by big data are the same as our observations on the behavior of 12 children. As for the gender preferences, there were no differences for different genders regardless of big data and children's behavior observations. This study found that: 1. There is a significant difference in age preference for games. 2. There is no significant difference in gender preference for games. Therefore, when we design game-based learning, we can use the results as a reference.

Keywords: Game-based learning, mobile game, a mixed study, age preference, gender preference.

1. Introduction

Nowadays, there are plenty of Game-based Learning games (Lin & Hou, 2016), especially educational mobile games for children (GSMA, 2015; Wishart, 2018). Children learn from play (Piaget, J. 1952), and children have their own preferences for games. Do children of different age groups or genders have different preferences for games? This research mainly used a mobile game as an example to explore children's preferences for the game by big data and behavior observation.

2. Research Method

This study adopted a mixed method to find out children's preference to the game. The mixed method of this study is based on big data and supplemented by behavior observation. The selected game for this study was Lily's Daily Life, which guides children to arrange life routine through role tasks. The way of playing is to allow children to enter different situations, such as going to school in the morning and going home in the evening. It lets children arrange life activities, and in the end, presents a picture of the story after children planned the situations. Please refer to Figure 1, Screenshot of Lily's Daily Life.

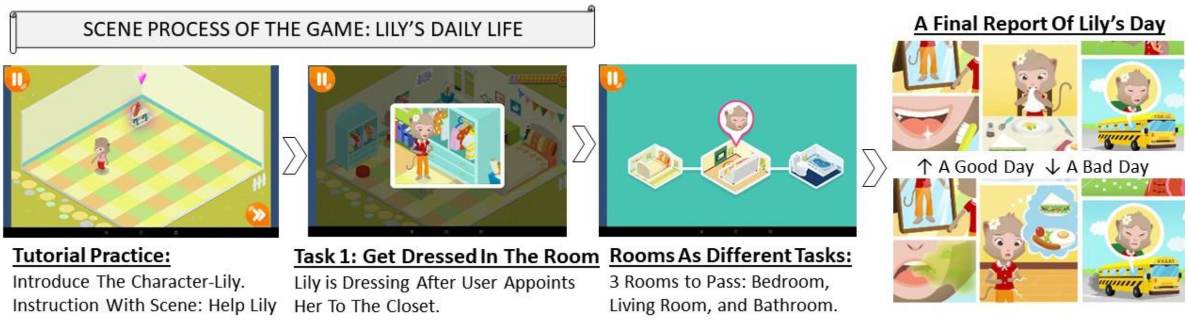
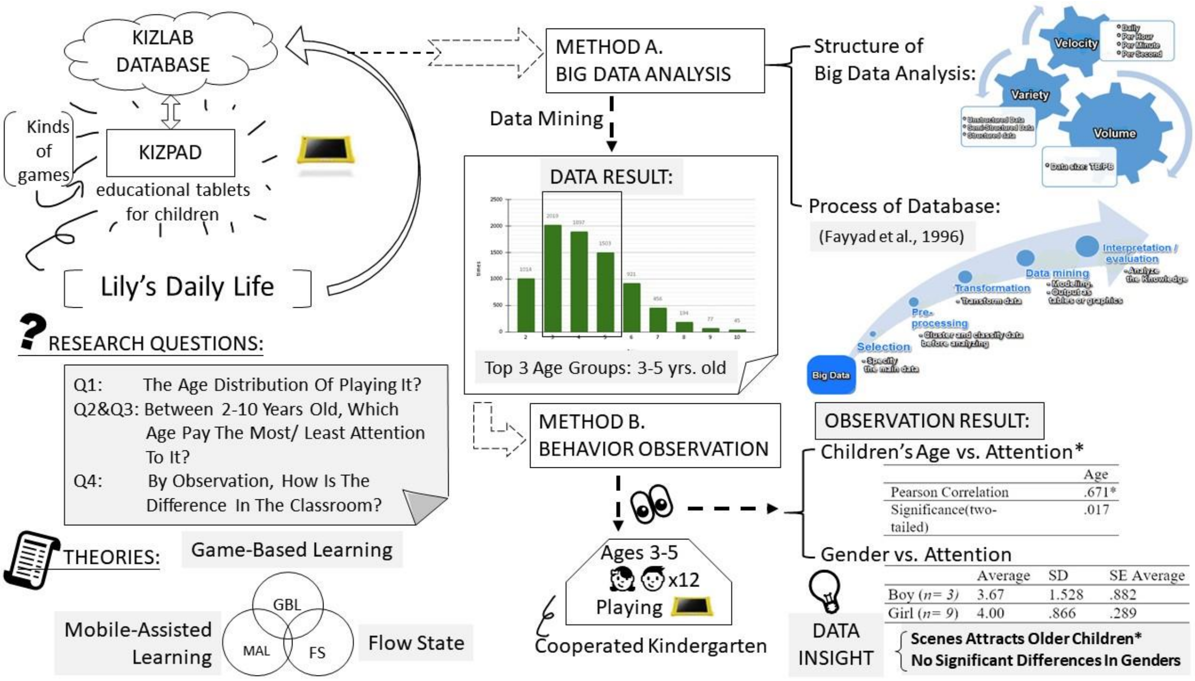

SCENE PROCESS OF THE GAME: LILY'S DAILY LIFE
A Final Report Of Lily's Day
↑ A Good Day ↓ A Bad Day
Tutorial Practice:
Introduce The Character-Lily.
Instruction With Scene: Help Lily
Task 1: Get Dressed In The Room
Lily is Dressing After User Appoints Her To The Closet.
Rooms As Different Tasks:
3 Rooms to Pass: Bedroom, Living Room, and Bathroom.

Figure 1. Screenshot of Lily's Daily Life.

3. Research Structure and Findings

The implementation steps of this research include: 1. Exploring children's preferences of different ages and genders in this game through big data. 2. In order to verify whether the big data information is true, we observed children's behaviors for this game in the classroom. The structure and findings of this research are shown in Figure 2.



Q1: The Age Distribution Of Playing It?
Q2&Q3: Between 2-10 Years Old, Which Age Pay The Most/ Least Attention To It?
Q4: By Observation, How Is The Difference In The Classroom?

	Age
Pearson Correlation	.671*
Significance(two-tailed)	.017

	Average	SD	SE Average
Boy (n= 3)	3.67	1.528	.882
Girl (n= 9)	4.00	.866	.289

Scenes Attracts Older Children*
No Significant Differences In Genders

Figure 2. Research Structure and Findings.

This study found: 1. Children of different ages have a preference for games. Take Lily's Daily Life as an example, children who prefer this game are 3-5 years old. 2. Children of different genders have no significant differences in game preferences. Taking Lily's Daily Life as an example, we found no significant differences.

The results of this study can be used as a reference for early childhood educators and game designers. This study only tested for one game, we will plan to discuss and compare more games in the future.

References

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37-37. https://doi.org/10.1609/aimag.v17i3.1230

GSMA. (2015). *Children's use of mobile phones: A special report 2014.* London and Tokyo: GSMA and NTT DOCOMO Inc. https://www.gsma.com/publicpolicy/wp-content/uploads/2016/09/GSMA2014_Report_ChildrensUseOfMobilePhonesASpecialReport.pdf

Lin, Y.-H. & Hou, H.-T. (2016). Exploring young children's performance on and acceptance of an educational scenario-based digital game for teaching route-planning strategies: a case study, *Interactive Learning Environments*, *24*(8), 1967-1980. https://doi.org/10.1080/10494820.2015.1073745

Piaget, J. (1952). Autobiography. In E. G. Boring, H. S. Langfeld, H. Werner, & B. M. Yerkes (Eds.), *A history of psychology in autobiography* (pp.237-256). Worcester, Clark University Press.

Wishart, J. (2018). *Mobile learning in schools: Key issues, opportunities, and ideas for practice.* Routledge.

The Impact of Augmented Reality on Vocabulary Learning of EFL Elementary School Students

Chin-Huang Daniel Liao[a*], Wen-Chi Vivian Wu[b], Chang Tin-Chang[c] & Chang-Hung Lee[d]
[a]*Department of Business Administration, Asia University, Taiwan*
[b]*Department of Foreign Languages and Literature, Asia University, Taiwan*
[c]*Department of Digital Media Design, Asia University, Taiwan*
[d]*Department of Electrical Engineering, National Sun Yat-sen University, Taiwan*
*danieliao820@gmail.com

Abstract: This study investigated the impact of the Augmented Reality (AR) on vocabulary learning of EFL students in Taiwan, the research site of this study. In this study, the experimental research design was employed, and multiple data sources were collected for data analysis, including a pretest, post-test, and on-site class observation. The experimental group received the treatment of the "AR word system of STEMUP", while the control group received the traditional teaching method. The class met for 40 minutes each week and the experiment lasted for one month. The research subjects of this study were the fifth-grade students of a primary school located in central Taiwan. In order to understand the effect of AR game-based instruction on students' English vocabulary learning, we selected single words (all of which are 1200 words by the Ministry of Education). Descriptive statistics, univariate analysis, and paired sample t-test were used STEMUP app system was applied in this study. Vocabulary AR will be presented in front of students by scanning the vocabulary picture to motivate student to learn English vocabulary. This study will provide directions on how AR game-based apps can be incorporated into language classrooms to enhance vocabulary learning of EFL learners.

Keywords: Vocabulary learning, augmented reality, english learning

1. Introduction

Vocabulary learning is a major factor in successful language acquisition for students who have English as a foreign language (EFL; Mearn, 1982). Elementary school is a very important stage for children to learn English. Many studies point out that a major activity in language learning is vocabulary acquisition, and vocabulary is key to the abilities of reading (Laufer & Paribakht,1998), speaking (Haynes & Baker, 1993), listening (Hincks, 2003), and writing(Hinkel, 2001). The most critical ability for students to improve their English learning is " vocabulary ability ". Students can read more books and understand more conversations. The process of learning English is like building a house and every word is like a brick. With a solid amount of words, students can learn English well. Principal Li Jiatong pointed out that "good English learning method is new words and grammar", and we can see the importance of words for learning English well. In addition, Vocabulary is considered essential to successful second/foreign language learning (Schmitt, 2000). Teachers must use various audio-visual media as learning aids to provide suitable learning content (Zhou, 1993).

The purpose of this study was to explore whether students' learning effectiveness and interest could be effectively improved by using the pretest, post-test analysis, and learning outcomes scale analysis. This study was to explore the effects of using an AR vocabulary learning app, as opposed to traditional vocabulary learning in the elementary school learning environment.

2. Literature Review

English Learning is very important in elementary school. Therefore, in addition to constructing the system of AR in English teaching, this study will also combine the use of tablet computers, which is expected to be more effective than hand-held mobile devices. Similarly, Santos et al. (2016) reported that the use of handheld AR could potentially lead to improved retention of words and keep students motivated and satisfied with vocabulary learning. Good visual display and complete information transmission can enhance learners' motivation and interest in English teaching and improve the quality and efficiency of English teaching. Therefore, the integration of AR into English courses can enhance students'interest in learning. Tsung-yu Liu and yuan-jen Chang two scholars also use the data from Taiwan as a basis to develop internationally recognized research. By using the experience and technology of these two scholars, we hope to develop AR learning system which is more suitable for children's English education.

3. Methodology

3.1 Research Design

Document analysis, interviews, and questionnaires were used in this study. The participants were fifth-grader students and divided into two groups in Taichung City. The experimental group was taught with an AR vocabulary app with speech input (StemUp). The Control group was acquired in the traditional teaching model. In each week, the children had a lesson of about 40 minutes to learn the relevant vocabulary. After the instruction, the two groups were once again assessed as to their knowledge of animal-related, stationary-related, and food-related vocabulary.

Both quantitative and qualitative methods were used in this study. Concerning students' questionnaires, were obtained for statistical analysis. The Pretest, The English reading and writing proficiency exam for these fifth-grader students, was used to identify students' English learning achievement.

3.2 Participants

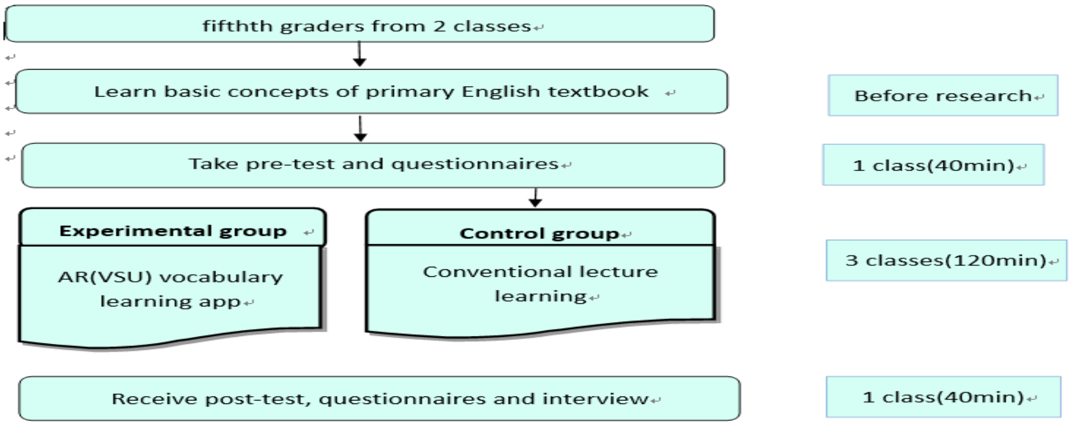
The participants were fifth-grader students and divided into two groups in Tanzi District, in Taichung City. The experimental group was taught with AR applied in vocabulary and pads were used. Control group was acquired in traditional teaching model with textbook.

3.3 StemUp System Framework and Function

StemUp (SU) is a set of the AR system, students with mobile phones or tablets, scanning AR marker, vivid display StemUp of the word, students can carry out three-dimensional learning. Students can play the game and answer the question correctly, will be awarded a gold coin, increase the motivation of students. In addition to reading the vocabulary, students can also practice pronunciation, students will be able to read the single word to pass. Some most important characteristics of StemUp are teachers can edit courses and create Multiplayer classrooms. In addition, the system can record the student's learning process, whether all of them moved, how much time it took to complete and how many times it took to answer correctly, and record the student's learning process in detail.

3.4 Experimental Procedure

The Experiment Procedure as shown in Figure 1:



fifthth graders from 2 classes
Learn basic concepts of primary English textbook
Take pre-test and questionnaires
Experimental group
AR(VSU) vocabulary learning app
Control group
Conventional lecture learning
Receive post-test, questionnaires and interview
Before research
1 class(40min)
3 classes(120min)
1 class(40min)

Figure 1. The Experiment Procedure

4. Results and Conclusion

From the literature review, most of the researches reveal that AR in English courses can increase students' motivation. Besides, AR courses could potentially lead to improved retention of vocabulary. Although this study has not been carried out because of covid-19 and students' summer vacation, some research outcomes could be used for future researchers. The teaching plan of the AR words of food, animal, and stationary in StemUp can be shared with researchers for AR-related research. In addition, researchers can download the StemUp in Andriod or ios app, please mail to me, danieliao820@gmail.com, and help you activate your account. We hope this AR system will be beneficial to researchers for future research.

5. References

Che Samihah Che Dalima, Mohd Shahrizal Sunar,Arindam Dey,Mark Billinghurst(2020). Using augmented reality with speech input for non-native children's language learning. *International Journal of Human-Computer Studies.134 (2020).44-64*

Ruo Wei Chen1 and Kan Kan Chan(2019). Using Augmented Reality Flashcards to Learn Vocabulary in Early Childhood Education. *Journal of Educational Computing Research* Volume: 57 issue: 7, page(s): 1812-1831. DOI: 10.1177/0735633119854028

Solak, E., & Cakir, R., (2016). Investigating the role of augmented reality technology in the language classroom. *Online Submission*, 18(4), 1067 – 1085.

Santos, M.E.C., Lubke, W., Taketomi, T., Yamamoto, G., Rodrigo, M.M.T., Sandor, C.,Kato, H., (2016). Augmented reality as multimedia : the case for situated vocabularylearning. Res. Pract. Technol. Enhanced earn. https://doi.org/10.1186/s41039-016-0028-2.

Boonbrahm, S., Kaewrat, C., Boonbrahm, P., (2015). Using Augmented Reality Technology in Assisting English Learning for Primary School Students. *Springer International Publishing*, Cham, pp. 24 – 32.

Wu, H. K., Lee, S. W. Y., Chang, H. Y., & Liang, J. C. (2013). Current status, opportunities and challenges of augmented reality in education. *Computers & education*, 62,41 – 49.

Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Integrating Parsons Puzzles with Scratch

Jeff BENDER[a*], Bingpu ZHAO[b], Lalitha MADDURI[a], Alex DZIENA[a], Alex LIEBESKIND[a] & Gail KAISER[a]

[a]*Programming Systems Laboratory, Columbia University, USA*
[b]*Department of Computer Science, Barnard College, USA*
*jeffrey.bender@columbia.edu

Abstract: We surveyed grade 6-9 teachers to learn teacher perceptions of student engagement with computational thinking (CT) and how well their needs are met by existing CT learning systems. The results and a literature review lead us to extend the trend of balancing Scratch's agency with structure to better serve learners and reduce burden on teachers aiming to learn and teach CT. In this paper, we integrate Parsons Programming Puzzles (PPPs) with Scratch and analyze the effects on adults, who crucially influence the education of their children. The results from our small pilot study suggest PPPs catalyze CT motivation, reduce extraneous cognitive load, and increase learning efficiency without jeopardizing performance on transfer tasks.

Keywords: computational thinking, parsons programming puzzles, scratch, motivation, cognitive load

1. Introduction

In response to a crisis in CS teacher certification and a deficit of student exposure to CS in grades K-12 (Wilson et al., 2010; Leyzberg et al., 2017), governments are enacting policies requiring CT in schools (Whitehouse.gov, 2016; The Royal Society, 2016). Additional argument (Wing, 2006, 2008) and evidence (Grover et al., 2013) provide rationale for ensuring children achieve CT competency during the formative cognitive and social development cycles throughout grades K-12. Parsons programming puzzles (PPPs), which enable learners to practice CT by assembling into correct order sets of mixed-up blocks that comprise samples of well-written code focused on individual concepts, are one approach used to introduce CT efficiently (Parsons et al., 2006; Ericson et al., 2018).

These scaffolded program construction tasks facilitate learning of syntactic and semantic language constructs underlying a CT concept. As the learner solves carefully designed single-solution puzzles, she arranges constructs from a curated assortment in a cycle of deliberate practice that exposes and addresses misconceptions (Kaczmarczyk et al., 2010; Emerson et al., 2020). Among the correct code fragments, she might find distractors which provoke cognitive conflict that reinforces learning (Karavirta et al., 2012). The partial suboptimal path distractor type, for example, might tempt a learner toward faulty progress without enabling her to solve the problem fully, thereby triggering recognition of a misconception and productive backtracking toward the correct solution (Harms et al., 2016).

Research indicates this structured approach to learning CT can lead to more efficient concept learning than alternatives such as learning by tutorial, or writing/fixing code (Harms et al., 2015; Ericson et al., 2017; Zhi et al., 2019). To measure efficiency, researchers often leverage cognitive load theory, which helps to distinguish between the complexity of the material, the instructional design, and the strategies used for knowledge construction. Since PPPs provide constrained problem spaces, they can induce lower cognitive load than that experienced when writing code with open-ended agency.

In the current study, we seek further evidence of their efficiency by integrating PPPs into Scratch, a block-based environment initially designed for informal learning that invites exploration, collaboration, and knowledge construction through personally meaningful creation (Maloney et al., 2010). K-8 teachers use Scratch more than any other coding language internationally (Rich et al., 2019), resulting in an ecosystem with over 74 million users (MIT Media Lab, 2021), and more research focus than any other environment in K-12 from 2012-2018 (McGill et al., 2020). However, historical findings indicate Scratchers infrequently demonstrate skill increases over time (Scaffidi et al., 2012), misconceive loops, variables, Booleans, nested conditionals, and procedures (Grover et al., 2017, 2018), and often adopt habits unaligned with accepted CS practice (Meerbaum-Salant et al., 2011). In a recent study of 74,830 Scratch projects, 45% contained at least one bug pattern (Frädrich et al., 2020). Instead of problem-solving algorithmically, Scratchers often engage in bricolage (Harel et al., 1991), which involves bottom-up tinkering that does not necessarily prove productive (Dong et al., 2019).

To balance this agency with structure as recommended in (Brennan, 2013), and to encourage the development of desired habits when learning CT concepts without stifling learner creativity, researchers have designed external Scratch curricula (Brennan et al., 2014; Franklin et al., 2020), created introductory Scratch Microworlds with reduced functionality (Tsur et al., 2018), and devised learning strategies based on the Use->Modify->Create pedagogy to scaffold instruction (Salac et al., 2020). We extend this trend by integrating with Scratch PPPs with explicit goals that offer gameful scoring targets and per-block feedback to disincentivize trial-and-error behavior and steer learners toward correct solutions. We reason that if learners initially can internalize CT concepts efficiently via PPPs, they can better deepen their understanding in less-restrictive interest-driven projects such as those described in (Kong et al. 2020) that embrace Scratch's roots in constructionism (Brennan et al., 2014).

To test this reasoning, we ran a pilot study targeting adults, who comprise a general population that might not only benefit from learning CT, but who might most effectively mobilize the advancement of teaching and learning CT for all. We investigated the following research questions: **R1**) what are the effects on motivation and cognitive load when training occurs via: PPPs; PPPs with distractors (PPPDs); programming with access to all blocks and without feedback (limited-constraint-feedback or LCF)?; **R2**) what are the effects on learning efficiency for training via PPP, PPPD, and LCF? Although the 75-participant sample limits the number of statistically significant results, findings indicate: **F1**) participants self-report higher motivation when training via PPPs and PPPDs, and less extraneous cognitive load when training via PPPs than via PPPDs or via LCF; **F2**) participants training via PPPs and PPPDs experience increased learning efficiency compared with those training via LCF.

We first review the background and the required software development. We then document the study purpose, formative and summative evaluations, and results before previewing future work.

2. Background

Since PPPs emerged in the CS literature as a new form of program completion problem in 2006, the community has investigated their strengths and weaknesses. Strengths include: scaffolded support of syntax and semantics learning; solvers with prior experience perform better and need less time (Harms et al., 2016); quicker grading and less grading variability than code writing problems (Ericson et al., 2017); easier detection of learning differences between students compared to code writing and code fixing problems (Morrison et al., 2016); a moderate correlation between PPP proficiency and code writing proficiency in an exam setting (Denny et al., 2008); less completion time required than for code writing exercises with equivalent performance on transfer tasks (Ericson et al., 2017; Zhi et al., 2019); higher enjoyment and less completion time required than for tutorial users with better performance on transfer tasks (Harms et al., 2016); and a lack of significant differences in performance across gender. Weaknesses include: constriction of puzzle-design surface to maintain single-solution structure (not strictly required, but commonly enforced to maintain strengths); the invitation of trial-and-error behavior in PPPs with excessive corrective feedback (Helminen et al., 2013); and a potential ceiling effect when feedback guides most learners to solve PPPs correctly, resulting in the need to evaluate learner process in addition to learner product when assessing (Helminen et al., 2012).

The community also has explored differences in learning outcomes resulting from using different PPP elements. Evidence suggests that 2D puzzles, in which the student must not only correctly order programming constructs but also indent them correctly, are more difficult than 1D (Ihantola et al., 2011). Similarly, PPPs that conceal the number of lines of code needed for each solution section and those that include distractors are more difficult, require more time to complete, and produce higher cognitive load during training than those that specify section sizes and those without distractors (Garner, 2007; Harms et al., 2015). Learning differences continue to emerge when researchers vary these elements. For example, learners struggle more when distractors are randomly distributed among the correct code constructs than when they are paired with correct constructs (Denny et al., 2008).

To identify these strengths, weaknesses, and learning differences between PPP elements, researchers often leverage Cognitive Load Theory (CLT) (Sweller, 2010). According to CLT, the brain provides limited short-term memory and processing capability along with infinite long-term memory, and learning occurs via schema construction and elaboration that leads to automation. Construction ensues by combining new, single elements into one larger element, and elaboration follows by adding new elements to an existing, larger element. Through intensive practice, individuals can automate their processing of these larger elements so that they execute without controlled processing.

CLT helps distinguish characteristics of and between PPP systems by offering a framework with tools to measure the three types of cognitive load experienced: intrinsic, extraneous, and germane. The total number of interacting elements perceived by the learner determines intrinsic load; the sometimes-impeding organization and presentation of the content determines extraneous load; and the instructional features necessary for schema construction determine the germane load. PPP designers aim to reduce extraneous load to free learners' capacity to contend with germane load when attempting to maximize learning efficiency. For example, the pairing of distractors with correct constructs might increase germane load by focusing student attention on the intended, misconception-revealing differences between two solution options, while also reducing extraneous load by eliminating the need to search for and identify the two relevant options amidst a random distribution of constructs.

To measure relative learning efficiency quantitatively across conditions, researchers calculate instructional and performance efficiency (van Gog et al., 2008). These calculations account for learners who compensate for an increase in mental load by committing more mental effort, thereby maintaining constant performance while load varies. The data recorded often include empirical estimates of mental effort during instruction (EI) and transfer (ET) tasks and the performance (P) on transfer tasks. The EI and P calculation measures the instructional efficiency of the learning process, while the ET and P calculation measures the performance efficiency of the learning outcome. For example, in a study that included interactive puzzles in the transfer phase, results indicate PPPs with randomly distributed distractors decrease performance efficiency (Harms et al., 2016). In our study, we measure instructional efficiency with a focus on learning process economy.

3. Software Development

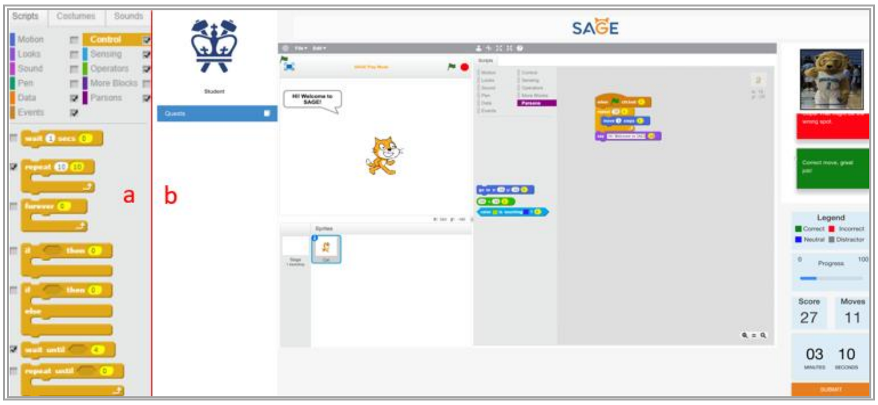


Figure 1. Design, Play, and Assessment Functionality integrated via PPP in Scratch.

To investigate our research questions, we modified Scratch to facilitate the design, play, and assessment of PPPs. Aligned with the gamification strategy described in (Tahir, Mitrovic, & Sotardi, 2020), in which the game elements were added to SQL-Tutor, and similar to recent iSnap integrations offering progress panels and adaptive messages (Zhi et al., 2019; Marwan et al., 2020), we augmented Scratch to influence the behavior of learners. As shown in Figure 1a, we first established a design mode which enables content developers to assign points to individual blocks and select blocks for inclusion in a new PPP palette. Equipped with this functionality, teachers can assign higher point values to blocks relevant to the CT concept studied and can isolate in a single palette blocks pertinent to the puzzle.

As presented in Figure 1b, we next established a play mode which enables students to load PPPs in a manner that displays the designed animated elements in the Scratch stage, but none of the blocks in the scripts pane authored as the solution. Technical detail is reported in (Sulaiman et al., 2019), but relevant to this study is an assessment system that includes a gameful scoring algorithm intended to encourage deliberate practice and discourage trial-and-error behavior. The longest common subsequence feedback algorithm described in (Karavirta et al., 2012) inspired this approach; ours differs in that we leverage block points, use them and subsequence length as multipliers, and sum the multiples from all subsequences matching the single correct solution while also deducting for incorrectness in absolute position. The closer the participant is to the solution, the higher the score.

Lastly, we added auto-initialization and auto-execution to reflect progress visually after each block placement during puzzle play. These mechanisms enable the display of gameful animations while an avatar presents per-block correctness feedback. They also calculate completion progress so that the learner receives appropriate feedback when she correctly solves the puzzle or the allotted time expires.

4. Study Purpose

This extended functionality positioned us to fill gaps in existing research. One study purpose was to explore the adult-use of CT learning system functionality primarily designed for children. Recent research has: 1) found significant correlation of motivation and previous programming experience with self-efficacy and inclination toward a CS career in elementary students (Aivaloglou et al., 2019); 2) indicated drag-and-drop programming can increase three CS motivational factors in middle school (Bush et al., 2020); 3) suggested computing experiences prior to university can affect the world-image of computing habits, perceptions, and attitudes which enable or inhibit pathways into CS (Schulte et al., 2007); and 4) illuminated benefits of community commitment and a CS/CT focused ecosystem inclusive of the home and community (Cao et al., 2020; DeLyser, 2018). Since demographic factors can drive communal values, and perceptions of how computing fulfills those values can affect sense of belonging and student retention (Lewis et al., 2019), we measure adult motivation and cognitive load while probing for attitudinal change that might influence the CT inclination for participants' children.

A second purpose was to further identify PPP elements that optimize learning efficiency, since the behavior of programming environments can affect novices' learning (Karvelas et al., 2020). While many researchers have hypothesized (Denny et al., 2008) and less often produced evidence (Ericson et al., 2018) that PPPs can result in more efficient learning than alternatives such as writing or fixing code, recently some have attempted to measure the contributions of various PPP elements (Kumar, 2017, 2019a, 2019b; Sirkia, 2016). We measure PPP learning efficiency with and without distractors, while offering a comparison to programming with LCF. Derived from the literature, our hypotheses were: **H1**) PPP and PPPD training increase motivation and reduce extraneous cognitive load compared to training via programming with LCF; **H2**) PPP training yields highest learning efficiency.

5. Formative Evaluation

As an early step in a roadmap of studies intended to explore the efficacy of adding gameful systems to novice programming environments, and with an aim to reinforce construct validity, we engaged in a formative evaluation with grade 6-9 educators. Through design thinking activities, we advanced our learning design technique, similar to the approach described in (Kashmira & Mason, 2020). Our goals included: 1) identifying the CT concepts receiving focus; 2) eliciting the pedagogical needs of practicing teachers; 3) and refining puzzle and feedback systems. We focus discussion here on goal 1.

5.1 Participants

The participants included 21 teachers from learning organizations such as Girls Who Code and codeHER, and 17 from U.S. schools. 11% had taught with Scratch for at least 2-4 years, 63% for 6-18 months, and 26% had instructed with Scratch for less than 6 months. 34% of the teachers used Scratch for at least 51% of their curriculum, 29% used it for 26-50%, and 29% used Scratch for at least 11-15%.

5.2 CT Concept Engagement

(Ihantola et al., 2016) highlights the concerning status quo in which most studies in the field focus on a single institution and a single course, without validation by subsequent replicating research, leading to limited understanding of the reasons results occur. To contribute replication results, and to identify the CT concepts receiving focus, we distributed a survey that included a question from a survey previously distributed to K-9 teachers in five European countries (Mannila et al., 2014). This question asks teachers to respond with their perceptions of student engagement in nine facets of CT. Since we targeted a narrower set of teachers in the U.S., it is perhaps unsurprising that the results do not match the earlier international study, in which teachers reported their students most frequently use CT concepts related to data (e.g. analysis). However, we present this finding to reinforce the replication concerns raised, and to underscore the challenges the community faces when attempting to disseminate CT globally.

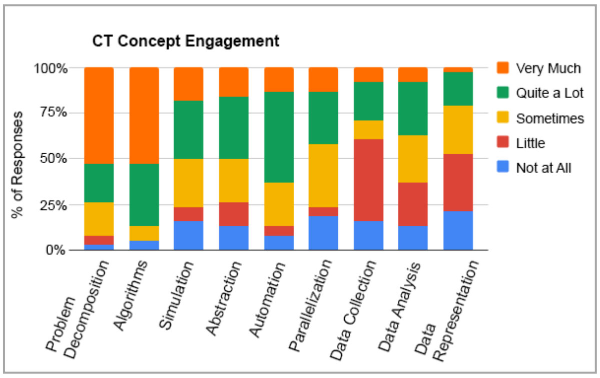


Figure 2. How a Small Sample of U.S. Teachers Perceive Student Engagement in CT Concepts.

Our findings in Figure 2 indicate teachers perceive their students engage in data CT concepts less than others such as abstraction and algorithms. Aside from the differences in population samples, and the associated threat to internal validity due to implicit differences in curricula (Barendsen et al., 2015) notes a low ratio of data knowledge in K-9 U.S. CSTA materials, 2%, compared with the English national curriculum, 14%, English Computing at School, 16%, and Italian guidelines, 25%), an extra explanation for this contrast could be related to the respondent recruitment process, as we specifically targeted Scratch teachers, whereas the earlier one did not. Since the small sample introduces a threat to external validity, future studies could try to replicate these results while controlling for technology and teacher pedagogical content knowledge utilizing a Content Representation approach like the one described in (Grgurina et al., 2014). Regardless, the lack of student engagement with data warrants investigation, as it is an alarming result for an increasingly data-driven society.

6. Summative Evaluation

6.1 Study Design

The formative evaluation helped us prioritize development, craft learning materials, and organize an initial summative evaluation via a 10-step between-subjects study through Amazon Mechanical Turk (Amazon, 2021). Participants used written instructions to guide their solving of four puzzles, and responded to a validated CS cognitive load component survey (CS CLCS) (Morrison et al., 2014), to a programming attitude Likert scale survey derived from categorized text-based responses by adults learners in (Charters et al., 2014), and to an intrinsic motivation Task Evaluation Questionnaire (TEQ) (SDT). Protocol materials are publicly available to facilitate replication at https://bit.ly/3uhSUd7.

We randomly assigned participants to one of three independent variable conditions: 1) PPP; 2) PPPD; 3) LCF. The dependent variables included time and performance on the pre/posttests, time and block moves in puzzles, and the cognitive load, programming attitude, and TEQ results.

6.2 Materials

We tested and refined our materials in collaboration with a high school teacher, 16 of her freshman physics students with little prior exposure to CT, and eight undergraduates with diverse majors. Tests included trials of the surveys and puzzles, and think-alouds in which the participant would interact with puzzles while verbalizing her thoughts. Results led to refinements such as puzzle theme modification, normalization of pre/posttest difficulty, and simplification of language used in survey questions.

6.3 Participants

Given Wing's mobilizing declaration that CT is a "fundamental skill for everyone, not just for computer scientists" (Wing, 2006), and with the interventionist spirit of design-based research (Barab, 2014), we sought a learner population that might not otherwise encounter an opportunity to engage with CT but might influence its trajectory in the lives of children. We recruited 75 adults with varying degrees (24% high school, 60% undergraduate, 16% graduate) and a variety of self-reported programming experience (low: 38; medium: 26; high: 11). 46 men and 29 women comprise the sample population sourced from eight countries including the U.S. (60%), India (20%), and Brazil (11%).

7. Analysis & Results

7.1 Data Collection & Processing

We created seven surveys in Qualtrics to capture data not directly collected by our CT learning system. To help measure performance and efficiency, we added instrumentation to: 1) record time from puzzle start until submission; 2) trace each block moved; and 3) calculate score. Since the data did not exhibit Shapiro-Wilk normality ($p<0.05$), we used non-parametric statistics, including Kruskal-Wallis and Pairwise Comparison of Condition tests between-subjects, and Wilcoxon tests within-subjects, to address skewness and kurtosis. We used guidelines for characterizing effect sizes in (Fritz et al., 2012).

7.2 Cognitive Load

We did not find significant differences in overall cognitive load during training between conditions. Upon analyzing subtypes, we found no notable differences in intrinsic and germane load, but moderate differences in extraneous load (PPP: M=3.12, SD= 3.26; PPPD: M=3.55, SD=3.62; LCF: M=3.90, SD=3.62). This result supports **H1**, as PPP participants self-reported lower extraneous load than PPPD participants, while LCF participants reported the highest. Since the LCF condition presented far more block choices and block search options than the PPPD condition, which in turn presented more choices than the PPP condition, this result indicates reducing impediments to block identification frees capacity for intrinsic and germane load. The higher extraneous cognitive load for training via PPPDs than with PPPs aligns with the findings in (Harms, et. al, 2016). Since pedagogically, distractors can challenge the learner to address misconceptions, we recommend further study to track cognitive load as agency increases, as misconceptions not addressed during structured learning could amplify in open-ended environments, resulting in higher cognitive load if measured in sum across a longitudinal span.

7.3 Performance

Though we did not find significant training performance differences across conditions, participants in the PPP and PPPD conditions interacted with the blocks significantly more with a relatively strong effect ($H(2)=21.14$, $p<0.001$, $\varepsilon^2=0.29$; PPP vs. LCF: $p=0.001$; PPPD vs. LCF: $p<0.001$). The fewer block moves made by participants in the LCF condition indicates some may have perceived the task as sufficiently overwhelming to decrease exploratory programming behavior probability.

In addition to analyzing aggregate puzzle performance, we compared individual puzzles. Participants in the PPP and PPPD conditions correctly solved puzzle 3 with a significantly higher score ($H(2)=18.44$, $p<0.001$) and executed more moves ($H(2)=14.772$, $p=0.001$) than those who trained with LCF. Since the solution to puzzle 3 involved 14 blocks, the second-highest count, this result suggests PPPs and PPPDs, which help learners focus on smaller block selection sets, might increase training performance and motivation as the difficulty of the puzzle increases.

Although participants in each condition solved more posttest than pretest questions correctly (PPP:M=0.65, SD=2.03; PPPD: M=0.82, SD=1.76, LCF: M=0.32, SD=2.25), with those in the PPPD condition yielding the highest increase, there is no significant difference on posttest performance across conditions (H(2)=1.335, p=0.513). This lack of transfer performance disparity between PPP and PPPD conditions ostensibly replicates findings in (Harms et al., 2016), and is similar to findings on PPP inter-problem and intra-problem adaptation in (Ericson et al., 2018).

7.4 Efficiency

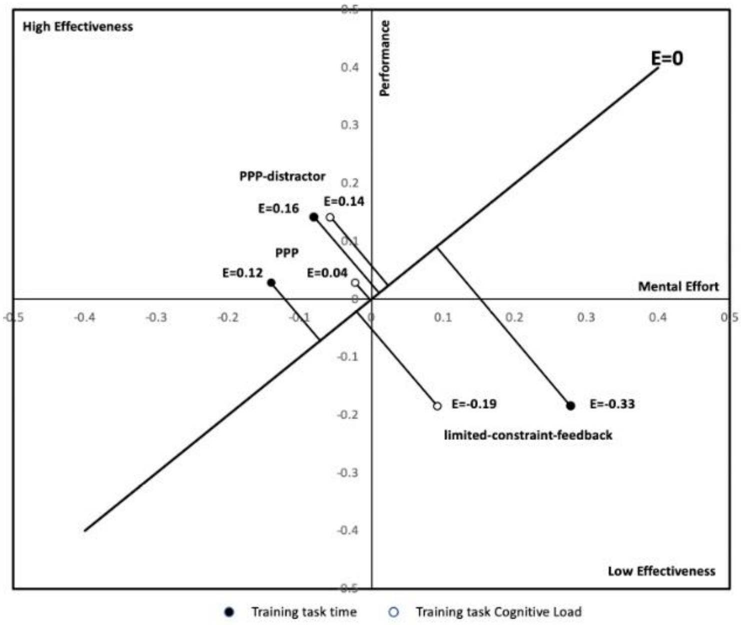


Figure 3. Instructional Efficiency (E) for each of the Three Conditions.

To measure efficiency, we analyzed training and transfer task time across conditions. During training, participants in the LCF condition, despite making fewer block moves, required significantly more time than those in the PPP and PPPD conditions with a moderate effect (H(2)=6.203, p=0.045, ε^2=0.08; PPP vs. LCF: p=0.030; PPPD vs. LCF: p=0.021). Since transfer task performance did not vary significantly across conditions, this result suggests training via PPPs and PPPDs enables more efficient CT learning. We did not, however, find a significant difference in the transfer task time (H(2)=0.883, p=0.643).

To emphasize the opportunity for efficient CT learning, we calculated instructional efficiency, using pre/posttest improvement to measure transfer performance and both time and cognitive load as measurements of mental effort during training (Pass & Merrienboer, 1993). Figure 3 presents areas of high and low effectiveness separated by the effort line E=0. The chart depicts higher instructional efficiency for training with PPPs and PPPDs than with LCF. However, this result refutes **H2**, as the PPPD condition yielded the highest instructional efficiency. This result contrasts with findings in (Harms et al., 2016), which found evidence of decreased learning efficiency from PPPDs, but it aligns with hypotheses regarding distractor learning benefits in (Parsons et al., 2006; Karavirta et al., 2012).

7.5 Motivation

To analyze motivation quantitatively, we scored the TEQ and calculated the within-subject change in programming attitude that occurred between the start and end of the study. Although there was no significant difference in TEQ results across conditions, for the perceived competence subscale, participants who trained with PPPs (M=4.89, SD=1.18) and PPPDs (M=4.91, SD=1.36) scored marginally higher than those who trained with LCF (M=4.53, SD=1.77).

We also found significant positive attitude changes from the start to end of the study. PPP participants' attitude shifted most by: *perceiving programming as more fun* with a small effect size (H(2)=2.392, p=0.017, r=0.07), *more enjoyable* with a medium effect (H(2)=2.392, p=0.017, r=0.44), *easier to start* with a large effect (H(2)=3.038, p=0.002, r=0.55), *less difficult to understand* with a large effect (H(2)=-3.343, p=0.01, r=-0.6), and *less of a foreign concept* with a large effect (H(2)=-3.074, p=0.002, r=-0.55). PPPD participants' attitudes also shifted positively by: *perceiving programming as easier to start* with a large effect (H(2)=2.514, p=0.012, r=0.54). LCF participants shifted least by: *perceive programming as more enjoyable* with a medium effect (H(2)=2.514, p=0.012, r=0.46). When including only those with little prior programming experience, PPP participants reported *programming more as something they want to learn* with a medium effect (H(2)=1.997, p=0.046, r=0.48) and *less boring* with a medium effect (H(2)=-1.961, p=0.050, r=0.48) in addition to the attitude shifts described above. Although these results indicate attitude improvement and support **H1**, the lack of longitudinal data poses a threat to internal validity, as we cannot claim change at study conclusion persists.

Table 1. *Within-subject Attitude Change. Positive shifts (p), negative shifts (n).* *p<0.05, **p<0.01

Programming is...	PPP	PPP-distractor	limited-constraint-feedback
something I've wanted to learn (p)	M=0.19, SD=1.40	M=0.27, SD=1.31	M=0, SD=1.19
fun (p)	M=0.74, SD=1.67*	M=0.40, SD=1.74	M=0.36, SD=1.43
Enjoyable (p)	M=0.90, SD=1.83*	M=-0.05, SD=1.68	M=0.68, SD=1.76*
important to know (p)	M=0.25, SD=1.48	M=-0.05, SD=1.17	M=0.09, SD=1.19
easy to start (p)	M=1.35, SD=2.29*	M=0.68, SD=1.13*	M=0.45, SD=1.71
something that takes practice (p)	M=0.065, SD=1.09	M=0.05, SD=1.29	M=-0.32, SD=1.17
too difficult to understand (n)	M-1.48=, SD=2.03**	M=-0.77, SD=1.77	M=-0.64, SD=1.89
boring (n)	M=-0.41, SD=1.6	M=-0.32, SD=1.17	M=-0.54, SD=1.90
a foreign concept (n)	M=-1.13, SD=1.83*	M=-0.27, SD=1.55	M=0, SD=2.07
too time consuming (n)	M=-0.35, SD=2.09	M=-0.09, SD=1.27	M=-0.09, SD=2.44

To supplement the quantitative results, we sought qualitative feedback by requesting that participants describe their attitude or view toward programming after the learning experience. For both those who self-reported low and high prior programming experience, we recorded more hesitant responses from those who trained via limited constraint and feedback than those who trained via PPPs and PPPDs. One LCF participant who selected "have tried programming activities, but have not taken a class" in the demographic survey, reflected on sustained struggle: "I still feel like programming is insanely complex. When I was in college I dropped out of computer science as soon as we started python. I just couldn't understand what we were doing, and maybe I could understand it if I really tried. It just seems to be better geared towards certain people." One PPP participant who recorded the same prior programming experience noted: "[T]his activity was somewhat easy but programming is really much harder than this. [B]ut this is a good way for a kid to start learning." One PPPD participant who selected "never attempted to program before" revealed potential for future pursuit of CT: "I would love to learn more about programming and encourage my son to start learning programming early." These results support **H1** and those in (Charters, Lee, Ko, & Loksa, 2014), which found significant attitude improvement regardless of gender and education level after a brief online programming experience.

8. Conclusion & Future Work

Our survey of grade 6-9 teachers exposed teacher perceptions of limited student engagement with data concepts central to CT. These results led us extend the trend of balancing Scratch's agency with structure to better serve learners and reduce burden on teachers. A small pilot study of an adult population using a learning system that integrates PPPs with Scratch yielded results indicating the structure provided by PPPs catalyzes motivation for CT, reduces extraneous cognitive load, and increases learning efficiency without sacrificing performance on transfer tasks.

While these results reveal opportunities to advance the teaching and learning of CT via augmentations to block-based programming environments, we remain cautious due to external validity limitations: the single CT concept, sequences, and small summative evaluation population (75 adults), threaten generalizability. In future work, we intend to study additional CT concepts, functionality variation, and participants, to identify factors supportive of reliably efficient and effective CT learning.

Acknowledgments

The Programming Systems Laboratory is supported in part by DARPA N6600121C4018, NSF CCF-1815494 and NSF CNS-1563555.

References

Aivaloglou, E., & Hermans, F. (2019). Early programming education and career orientation: the effects of gender, self-efficacy, motivation and stereotype. *ACM SIGCSE*, (pp. 679-685).

Amazon. (2021). Retrieved from Amazon Mechnical Turk: https://www.mturk.com/

Barab, S. (2014). Design-based resarch: a methodological toolkit for engineering change. In K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences.* Cambridge University Press.

Barendsen, E., Mannila, L., Demo, B., Grgurina, N., Izu, C., Mirolo, C., . . . Stupurienė, G. (2015). Concepts in K-9 computer science education. *ACM ITiCSE-WGR*, (pp. 85-116).

Brennan, K. (2013). *Best of both worlds: issues of structure and agency in computational creation*. MIT.

Brennan, K., Balch, C., & Chung, M. (2014). *Creative computing.* Harvard Graduate School of Education.

Bush, J. B., Gilmore, M. R., & Miller, S. B. (2020). Drag and drop programming experiences and equity: analysis of a large scale middle school student motivation survey. *ACM SIGCSE*, (pp. 664-670).

Cao, L., Rorrer, A., Pugalee, D., Maher, M. L., Dorodchi, M., Frye, D., . . . Wiebe, E. (2020). Work in progress report: a STEM ecosystem approach to CS/CT. *ACM SIGCSE*, (pp. 999-1004).

Charters, P., Lee, M., Ko, A., & Loksa, D. (2014). Challenging stereotypes and changing attitudes: the effect of a brief programming encounter on adults' attitudes toward programming. *ACM SIGCSE*, (pp. 653-658).

DeLyser, L. (2018). A community model of CSforALL. *ACM ITiCSE*, (pp. 99-104).

Denny, P., Luxton-Reilly, A., & Simon, B. (2008). Evaluating a new exam question: parsons problems. *ICCE.*

Dong, Y., Marwan, S., Catete, V., Price, T., & Barnes, T. (2019). Defining tinkering behavior in open-ended block-based programming assignments. *SIGCSE*, (pp. 1204-1210).

Emerson, A., Smith, A., Rodriguez, F., W. E., & Bradford, W. (2020). Cluster-based analysis of novice coding misconceptions in block-based programming. *ACM SIGCSE*, (pp. 825-832).

Ericson, B., …Rick, J. (2018). Evaluating the efficiency and effectiveness of adaptive parsons problems. *ICER.*

Ericson, B., … Rick, J. (2017). Solving parsons problems versus fixing and writing code. *Koli Calling CER.*

Frädrich, C., Obermüller, … Fraser, G. (2020). Common bugs in Scratch programs. *ACM ITiCSE*, (pp. 89-95).

Franklin, D., Weintrop, D., Palmer, J., Coenraad, M., Cobian, M., Beck, K., . . . Crenshaw, Z. (2020). Scratch Encore: the design and pilot of a culturally-relevant. *ACM SIGCSE*, (pp. 794-800).

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates. *Experimental Psychology, 141*(1), 2-18.

Garner, S. (2007). An exploration of how a technology-facilitated part-complete solution method supports the learning of computer programming. *Informing Science & Information Technology.*

Grgurina, N., Barendsen, E., Zwaneveld, B., van Veen, K., & Stoker, I. (2014). Computational thinking skills in dutch secondary education: exploring pedagogical content knowledge. *ACM Koli Calling*, (pp. 173-174).

Grover, S., & Basu, S. (2017). Measuring student learning in introductory block-based programming: Examining misconceptions of loops, variables, and boolean logic. *ACM SIGCSE*, (pp. 267-272).

Grover, S., & Pea, R. (2013). Computational thinking in K–12: a review of the state of the field. *Edu. Res., 42*(1).

Grover, S., Basu, S., & Schank, P. (2018). What we can learn about student learning from open-ended programming projects in middle school computer science. *ACM ITiCSE*, (pp. 999-1004).

Harel, I., & Papert, S. (1991). *Constructionism.* Ablex Publishing.

Harms, K., Balzuweit, E., Chen, J., & Kelleher, C. (2016). Learning programming from tutorials and code puzzles: children's perceptions of value. *IEEE Visual Languages and Human-Centric Computing*, (pp. 59-67).

Harms, K., Chen, J., & Kelleher, C. (2016). Distractors in parsons problems decrease learning efficiency for young novice programmers. *International Computing Education Research*, (pp. 241-250).

Harms, K., Rowlett, N., & Kelleher, C. (2015). Enabling independent learning of programming concepts through programming completion puzzles. *IEEE Visual Lanugages and HCC*, (pp. 271-279).

Helminen, J., … Alaoutinen, S. (2013). How do students solve parsons programming problems? *L. & T. in Comp*.

Helminen, J., … Malmi, L. (2012). How do students solve parsons programming problems? *ICCE.*

Ihantola, P., & Karavirta, V. (2011). Two-dimensional parson's puzzles *IT Education, 10*, 119-132.

Ihantola, P., Vihavainen, A., Ahadi, A., Butler, M., Börstler, J., Edwards, S., . . . Rubio, M. (2016). Educational data mining and learning analytics in programming: literature review and case studies. *ITiCSE WGR.*

Kaczmarczyk, & Petrick. (2010). Identifying student misconceptions of programming. *SIGCSE*, (pp. 107-111).

Karavirta, V., Helminen, J., & Ihantola, P. (2012). A mobile learning application for parsons problems with automatic feedback. *International Conference on Computing Education Research*, (pp. 11-18).

Karvelas, I., Li, A., & Becker, B. A. (2020). The effects of compilation mechanisms and error message presentation on novice programmer behavior. *ACM SIGCSE*, (pp. 759-765).

Kashmira, D., & Mason, J. (2020). Empowering learning designers through design thinking. *ICCE* (pp. 497-502).

Kong, S., Lai, M., & Siu, C. (2020). Development of CT concepts in Scratch programming. *ICCE*, (pp. 652-657).

Kumar. (2017). The effect of providing motivational support in parsons puzzle tutors. *AI in Ed.*, (pp. 528-531).

Kumar, A. N. (2019a). Helping students solve Parsons puzzles better. *ACM ITiCSE*, (pp. 65-70).

Kumar, A. N. (2019b). Mnemonic variable names in Parsons puzzles. *ACM CompEd*, (pp. 120-126).

Lewis, C., Bruno, P., Raygoza, J., & Wang, J. (2019). Alignment of goals and perceptions of computing predicts students' sense of belonging in computing. *ACM ITiCSE*, (pp. 11-19).

Leyzberg, D., & Moretti, C. (2017). Teaching CS to CS teachers. *ACM SIGCSE*, (pp. 369-374).

Maloney, J., .. Eastmond, E. (2010). The Scratch programming language and environment. *Computing Education*.

Mannila, L., Dagiene, V., Demo, B., Grgurina, N., Mirolo, C., Rolandsson, L., & Settle, A. (2014). Computational thinking in K-9 education. *Innovation & Technology in Computer Science Education.*

Marwan, S., Gao, G., Fisk, S., Price, T., & Barnes, T. (2020). Adaptive immediate feedback can improve novice programming engagement and intention to persist in computer science. *ACM ICER*, (pp. 194-203).

McGill, M. M., & Decker, A. (2020). Tools, languages, and environments used in primary and secondary computing education. *ACM ITiCSE*, (pp. 103-109).

Meerbaum-Salant, O., Armoni, M., & Ben-Ari, M. (2011). Habits of programming in Scratch. *ACM ITiCSE.*

MIT Media Lab. (2021). *Scratch Statisitcs*. Retrieved from https://scratch.mit.edu/statistics/

Morrison, B., Dorn, B., & Guzdial, M. (2014). Measuring cognitive load in introductory CS. *ICER*, (pp. 131-138).

Morrison, B., … Guzdial, M. (2016). Subgoals help students solve Parsons problems. *ACM SIGCSE*, (pp. 42-47).

Parsons, D., & Haden, P. (2006). Parson's programming puzzles: a fun and effective learning tool for first programming courses. *Australian Conference on Computing Education.*

Pass, F., & Merrienboer, J. (1993). The efficiency of instructional conditions. *Human Factors, 35*(4), 737-743.

Prolific. (2021). Retrieved from Prolific | Online participant recruitment: https://www.prolific.co/

Rich, P. J., Browning, S., Perkins, M., Shoop, T. Y., & Belikov, O. M. (2019). Coding in K-8: international trends in teaching elementary/primary computing. *Tech Trends*, *63*(3), 311-329.

Salac, J., Thomas, C., Butler, C., Sanchez, A., & Franklin, D. (2020). TIPP&SEE: a learning strategy to guide students through use-modify Scratch activities. *SIGCSE*, (pp. 79-85).

Scaffidi, C., & Chambers, C. (2012). Skill progression demonstrated by users in the Scratch animation environment. *International journal of Human-Computer Interaction, 28*, 383-398.

Schulte, C., & Knobelsdorf, M. (2007). Attitudes towards computer science-computing experiences as a starting point and barrier to computer science. *ACM ICER*, (pp. 27-38).

SDT. (n.d.). Self-determination Theory: https://selfdeterminationtheory.org/intrinsic-motivation-inventory/

Sirkia, T. (2016). Combining parson's problems with program visualization in CS1 context. *Koli Calling*.

Sulaiman, J., Dziena, A. B., & Kaiser, G. (2019). SAGE-RA: a reference architecture to advance the teaching and learning of computational thinking. *Embedding AI in Education Policy and Practice for Southeast Asia.*

Sweller, J. (2010). *Cognitive Load Theory: Recent Theoretical Advances.* Cambridge University Press.

Tahir, F., Mitrovic, A., & Sotardi, V. (2020). Investigating the effects of gamifying SQL-Tutor. *ICCE*.

The Royal Society. (2016). *After the Reboot: Computing Ed. in UK Schools*. Retrieved from https://royalsociety.org/~/media/policy/projects/computing-education/computing-education-report.pdf

Tsur, M., & Rusk., N. (2018). Scratch microworlds: designing project-based introductions to coding. *SIGCSE*.

van Gog, T., & Paas, F. (2008). Instructional efficiency. *Educational Psychologist*, (pp. 16-26).

Whitehouse.gov. (2016). *CS for All* https://obamawhitehouse.archives.gov/blog/2016/01/30/computer-science-all

Wilson, C., Sudol, L., Stephenson, C., & Stehlik, M. (2010). *Running on Empty: The Failure to Teach K-12 CS in the Digital Age.* CSTA. http://www.acm.org/runningonempty/fullreport2.pdf

Wing, J. (2006). Computational Thinking. *Communications of the ACM*, *49*(3), 33-35.

Wing, J. (2008). Computational thinking and thinking about computing. *Philosophical Transactions.*

Zhi, R., Chi, M., Barnes, T., & Price, T. W. (2019). Evaluating the effectiveness of parsons problems for block-based programming. *ACM ICER*, (pp. 51-59).

Tactical Knowledge Acquisition Support System from Play Videos of Esports Experts

Minato SHIKATA[a*] & Tomoko KOJIRI[b]
[a]*Graduate School of Science and Engineering, Kansai University, Japan*
[b]*Faculty of Engineering Science, Kansai University, Japan*
*k667670@kansai-u.ac.jp

Abstract: Esports are a competition that involves computer games and video games that are played online. Some expert players upload videos of their play (play video) on the internet. In esports, understanding various tactical knowledge is important so that players can take appropriate actions based on such situations. Unfortunately, players have inadequate opportunity to learn them. The video performances of experts who play esports contain scenes where various tactical knowledge has been adopted by the experts. The objective of our research is to support players' grasping about the tactical knowledge of experts from their play videos. This research provides a system that extracts such components of tactical knowledge as sequence of actions, applicable situations, and its effects from good scenes that are selected by players from the play videos.

Keywords: computer game, observational learning, play videos, tactics acquisition

1. Introduction

Esports, which is an abbreviation of "electric sports," denotes competitions of computer games and video games. They are originated in a small competition at Stanford University in the United States in 1972. Their play popularity continues to rise as well as the numbers of people who engage in them (Jenny et al., 2017). Various types of games are provided as esports, and the skills required to play them are based on particular games. However, in all esports, players need to choose actions and strategies and exploit decision-making abilities to take appropriate actions that react to a competition's situation. This is the same as physical sports. That is, a soccer requires the skills of kicking the ball and a tennis requires the skills of hitting the ball, but decision-making abilities are just as important for both competitions.

Baker et al. argued that since decision-making abilities in physical sports such as soccer and tennis are implicit and gained through experience (Baker et al., 2003), few effective training methods have been established. In physical sports, players usually learn the appropriate actions for each situation from constant feedback and comments from coaches, managers, and teammates when reflecting their play. Unfortunately, players who do not share the same physical space with such experts lack opportunities to be given comments.

To compensate for the absence of experts and provide an opportunity to learn from them, Zhai et al. created an online coaching-support system for esports (Zhai et al., 2005). In esports, plays are easily recorded as videos (play video) and are shared through the Internet. In the system of Zhai et al., an expert and players in a distributed environment can interact and exchange annotated snapshots extracted from play videos. Based on annotations, the scenes where coaches provide comments can be grasped. However, since the quality of the comments differs based on the coaches, players are not always able to understand which actions are appropriate.

Some systems foster decision-making skills. Takahashi et al. proposed a training system for decision-making in disasters (Takahashi et al., 2017). Their system contained branched-training scenarios that are created to cope with unexpected situations during a disaster and simulates them with VR images. At branch points, users need to make decisions whose results are provided as selected scenes. Since the knowledge that should be acquired in a disaster can be assumed in advance, scenarios can be prepared. However, in esports with a short history, since the target decisions to learn are not clear, creating scenarios is difficult. For making appropriate decisions, understanding effective actions as tactical knowledge and applying them based on the situation are important.

In the field of physical sports, some systems provide analytical tools to grasp the tactical knowledge (Saito et al., 2015, Wu et al., 2017). These systems only provide the analysis tools, such as to derive the characteristics of winning games, but the way of interpreting the results is not supported. Thus, players without analytical skills cannot acquire tactical knowledge by themselves. In order to support players to acquire tactical knowledge independently, the system needs to extract tactical knowledge automatically.

Some systems disseminate tactical knowledge by giving feedback about players' actions (Vales-Alonso et al., 2015, Janusz et al., 2018). Feedback emphasizes inappropriate actions and suggests alternative strategies. Both of these research only suggest appropriate actions to the current situation without indicating the appropriate situations in which to apply them. Therefore, players struggle to apply the actions to other situations. The effect of applying the suggested actions is also unclear.

The objective of this research is to support players who want to acquire tactical knowledge that consists of a sequence of actions, the situations in which to apply it, and its effect. The moves of players are regarded as scenes in which tactical knowledge is being applied. Novice players must grasp the tactical knowledge from the plays and decisions of other players, especially from better, more experienced players. In esports, videos of the games played by good players can be easily obtained through the Internet. This research develops a tactical knowledge acquisition support system that extracts such components of tactical knowledge as sequences of actions, situations in which to apply them, and their effects from scenes selected by players from play videos.

2. Overview of Tactical Knowledge Acquisition Support System

2.1 Decision Making in Esports

In ball games such as tennis and soccer, in which the ball is used as an intermediary to alternately interfere with the opponent and compete directly, Nakagawa defined decision making as "making a decision about the play to be performed in the game" and argued that it consists of the four steps shown in Figure 1 (Nakagawa, 1984). The decision-making process in esports takes these four steps.

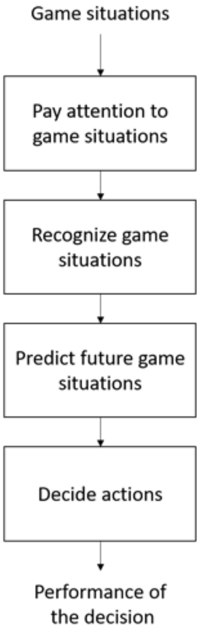

Game situations
Pay attention to game situations
Recognize game situations
Predict future game situations
Decide actions
Performance of the decision

Figure 1. Decision-making Process in Ball Games (Nakagawa, 1984).

In the “pay attention to game situations,” players concentrate on the appropriate elements of the perceived situation. In “recognize game situations,” they evaluate the perceived elements to react to the current game situation. “Predict future game situations” infers the future game situations based on recognized current game situations and action candidates. Since possible future situations differ depending on the candidates of the sequence of actions, various situations can be inferred. In “decide actions,” players determine potential strategies based on those that the player can exploit to her advantage. To select an effective action for a situation, future situations need to be weighed at the “predict future game situations.” Since future situations are derived by applying applicable actions to the situation, recognizing applicable actions from the situation is important. This paper defines the possible actions from each situation as tactical knowledge and helps players acquire such knowledge.

2.2 Tactical Knowledge Acquisition Support System

This paper defines tactical knowledge consists of a sequence of actions, situations where those actions can be applied, and their effects. For example, in baseball, when the out count is 0 and the player hits a fly ball to the outfield, the runner on the third base runs to the home base. In this tactical knowledge of the runner, a sequence of actions is “run and touch the home base”, and the situation of applying this tactical knowledge is “out count is 0 and the player hits a fly ball.” Its effect is “to gets 1 point.”

Players apply tactical knowledge during games. Play videos indicate scenes where such knowledge is applied and the application results. Therefore, a tactical knowledge acquisition support system encourages the comprehension of such knowledge from the play videos of experts.

If tactical knowledge is effective, a player’s situation becomes more advantageous than her opponent. Therefore, the sequence of actions that changes her situation in her favor is regarded as effective tactical knowledge for her. For example, in fighting games, if a player successfully decreased the life (HP gauge) of her opponent, her actions are labeled as successful tactical knowledge for that action that decreased the HP gauge.

Applying actions is one situation where tactical knowledge can work well and improve a player’s outcome. However, a situation is comprised of various elements, all of which are not always necessary for applying tactical knowledge. Some elements are relevant to the application of tactical knowledge; some are not. Relevant elements may commonly be recognized in scenes where identical tactical knowledge is applied. Our system provides an environment that allows users to cut out scenes from play videos to support their acquisition of tactical knowledge by presenting actions, situations, and effects from the common points of the extracted scenes.

Figure 2 shows an overview of our system. The interface shows the play videos and provides functions for extracting scenes deemed useful by users. For supporting the selection of scenes, the interface provides a time chart that shows the changes of the elements of the situations and the actions of the players who changed them. Annotations are attached to the extracted scenes that represent good types that are stored as candidates of tactical scenes. Annotations can be freely attached by users, for example, “a scene that decreases my opponent’s HP gauge” or “a scene that forces my opponent to another room.”

The interface for extracting tactical knowledge helps users acquire it from the extracted scenes. A list of the annotations of extracted scenes is presented. Based on the selection of annotations by users, a function, which extracts tactical knowledge, extracts sequences of meaningful actions and meaningful elements of the start situations of the scenes with identical annotations as the tactical knowledge in scenes.

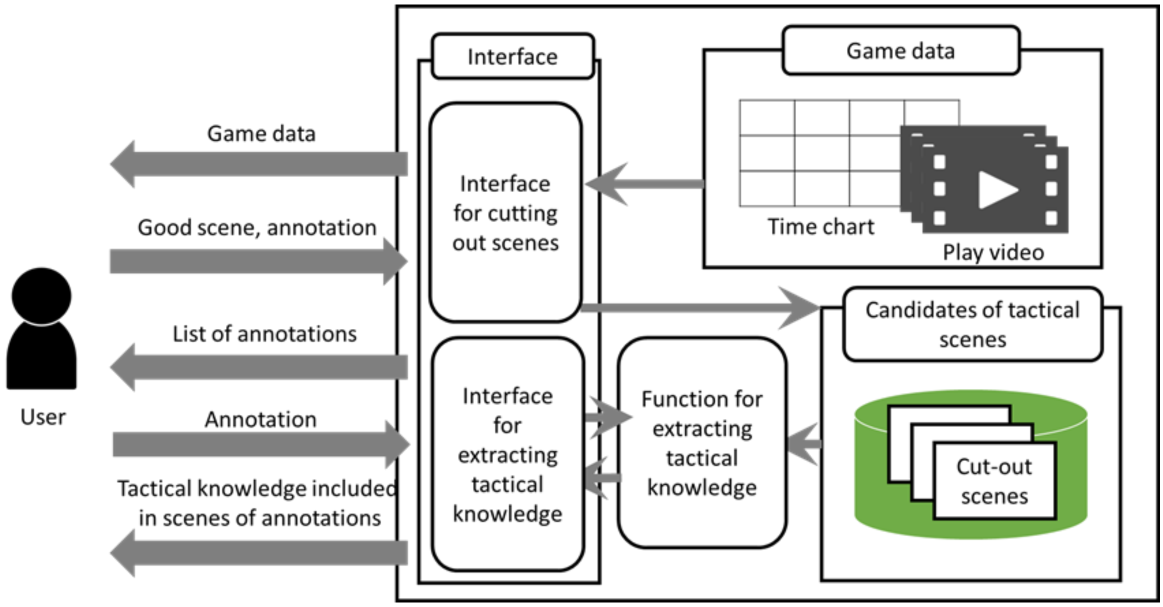

Interface
Game data
Game data
Interface for cutting out scenes
Time chart
Play video
Good scene, annotation
Candidates of tactical scenes
List of annotations
User
Annotation
Interface for extracting tactical knowledge
Function for extracting tactical knowledge
Cut-out scenes
Tactical knowledge included in scenes of annotations

Figure 2. System Overview.

3. Data Representation

3.1 Time Chart

A time chart expresses the game's progress by the actions taken in one time step and the change of the situation's elements based on such actions.

As an example, Table 1 shows part of the time chart for a fighting game, called Street Fighter V. In this game, there are two characters: the one operated by the player who is making the video and his opponent. They attack each other until either of them loses his life or the time is up. The first two lines show the actions for each time step. Since both players usually take actions simultaneously, the actions of two characters are written on one row. From the third line, states are shown. Their states are represented by the position, the remaining life (HP gauge), the power to use the strong attacks (EX gauge), and the power to use the special technique (V gauge). The position is set in either the field or near a wall. An HP gauge is represented by four values: high, medium, low, and minimal. EX and V gauges are expressed by 0, 1, 2, and 3. The state of the game space shows the positional relations between two characters and the remaining time. The positional relation is represented by three values: far, near, and close. The remaining time is represented as much or little.

Currently the time charts are prepared manually by authors. To create them automatically by analyzing the play videos remains as future work.

Table 1. *Part of Time Chart for Street Fighter V*

Action	Player's character		Kick	Fire ball	Fire ball	Fire ball
	Opponent's character		Defend	Bending backward	Defend	Fire ball
State	Player's character	Position	Field	Field	Field	Field
		HP gauge	High	High	High	High
		EX gauge	2	2	2	2
		V gauge	2	2	2	2
	Opponent's character	Position	Field	Field	Field	Field
		HP gauge	Low	Low	Low	Low
		EX gauge	1	1	1	1
		V gauge	2	2	2	2

Game space	Positional relation of two characters	Close	Close	Far	Far
	Remaining time	Much	Much	Much	Much

3.2 Tactical Scene

Tactical scenes are stored by scene data and annotations are attached by users. Scene data consist of five elements: the actions of the player's character, the actions of the opponent's character, the start situation, the end situation, and effects that consist of differences between the start and end situations. The player and opponent actions are the sequence of actions from the start to the end of the scenes. The situation consists of the values of each state element. These four elements are automatically generated based on the time chart when the scene was extracted by the user.

4. Function for Extracting Tactical Knowledge

The function that extracts tactical knowledge, such as meaningful actions, the elements of situations, and effects, from candidates of tactical scenes of identical annotation.

The meaningful start situations are the common elements of the start situations, and the meaningful end situations are the common elements of the end situations of the candidates of tactical scenes of identical annotation. The effect is the difference between the meaningful start and end situations. Table 2 shows example of start situations of tactical scenes 1 and 2. In this case, the meaningful elements are: Player.Position=Field, Opponent.Position=Field, Opponent.HP=Low, Opponent.EX=1, Game space.Remaining time=Much.

Table 2. *Example of Start Situations of Tactical Scene 1 and 2*

State		Tactical scene 1	Tactical scene 2
Player's character	Position	Field	Field
	HP gauge	High	Medium
	EX gauge	3	1
	V gauge	3	2
Opponent's character	Position	Field	Field
	HP gauge	Low	Low
	EX gauge	1	1
	V gauge	2	3
Game space	Positional relation of two characters	Close	Far
	Remaining time	Much	Much

On the other hand, a meaningful sequence of actions is also a common sequence of the actions of the candidates of tactical scenes. Since various sequences of actions can cause the same effect, the candidates of tactical scenes should be classified into groups to derive the common sequences of actions within groups. To classify the candidates of tactical scenes, the similarity between action sequences must be calculated. In this study, we applied the Levenshtein distance and defined the number of operations to make one action sequence to the other as a similarity. Here the operation is comprised of insert, delete, or replace. For example, in Figure 3, since one action is deleted from sequence 1 to become sequence 2, their similarity is 1. Tactical scenes whose similarity is small are classified as one group.

However, since time charts have every action for each time step, not all actions are important. For example, assume that throwing fire balls after kicking is critical tactical knowledge, and a small movement before throwing the fire balls does not affect the tactical knowledge. In this case, adding or deleting the movement action between a kick and a fire ball is meaningless. Since such meaningless operations do not affect the calculation of similarity, this study assigns their values as 0 for calculating similarity.

Meaningless operations differ based on the game. For example, Street Fighter V has nine action categories: punch, kick, throw, launching a projectile, counter, special move, defensive move, movement, and jump. The following are regarded as meaningless operations:

1. inserting/deleting/replacing a movement action;
2. replacing actions of the same category;
3. inserting/deleting actions of the same category for both sequences.

According to this definition, the similarity of action sequences in Fig. 3 is 0, since the value of deleting the movement action is 0.

By calculating the similarity, the candidates of tactical scenes whose similarity is zero form one group and the action's minimal sequence is regarded as the action sequence of the tactical knowledge.

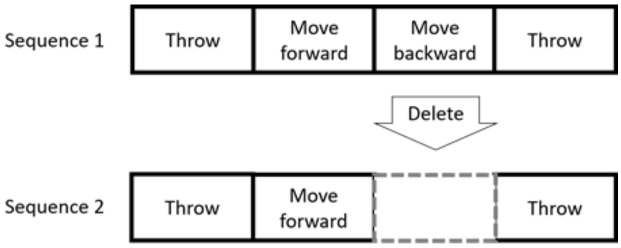


Figure 3. Example of Action Sequences.

5. Prototype System

We implemented our tactical knowledge acquisition support system in C # programming language. In this system, a user has only two main operations to view the tactical knowledge. The first is to cut out some scenes that they think are good for games. The second is to select what a user wants to view from the tactical knowledge automatically extracted by the system.

When it starts, a home screen appears that consists of two tabs. When a user selects a play video from the list and presses the OK button, she moves to the interface for extracting scenes. When she presses the tactical knowledge extraction button, she moves to the interface to extract tactical knowledge.

Figure 4 shows the interface for cutting scenes. A selected play video is shown in the displaying play video area and its time chart is shown on the displaying time chart area. When a user clicks on a certain point in the area for displaying time charts, the play video's seek bar moves to the corresponding scene, where the user can watch it. The user presses the mode change button to shift to the scene cutout mode from the play video watching mode. Here the user can extract scenes by selecting the start and end scenes and pushing the decision button. The sequence of actions in the cut-out scene is displayed in a list that displays action sequences. The left side of the action sequence list is a combination box for annotations, where the user can attach them to cut-out scenes. When the user presses the mode changing button again, the scene cutout mode ends, and the mode returns to the play video watching mode. By pushing the save button, the system stores the extracted scenes in an XML format.

Figure 5 shows the interface for extracting tactical knowledge. When a user selects from a list of annotations, a list of action sequences of the tactical scenes of the selected annotations is displayed in the list of the action sequences among all the tactical scenes. When a user selects from the list of meaningless operations, the similarity is calculated based on the selection methods, and meaningful actions are shown in the list of meaningful action sequences. If a user selects a meaningful action sequence, its start situations and effects are displayed in the displaying tactical knowledge area.

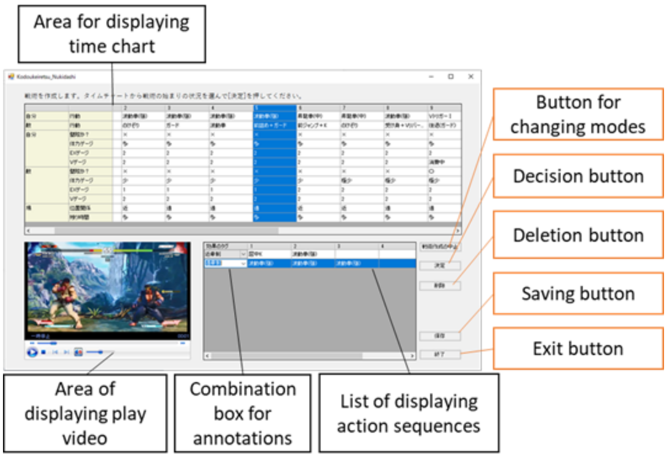

Area for displaying time chart
Button for changing modes
Decision button
Deletion button
Saving button
Exit button
Area of displaying play video
Combination box for annotations
List of displaying action sequences

Figure 4. Interface for Cutting out Scenes.

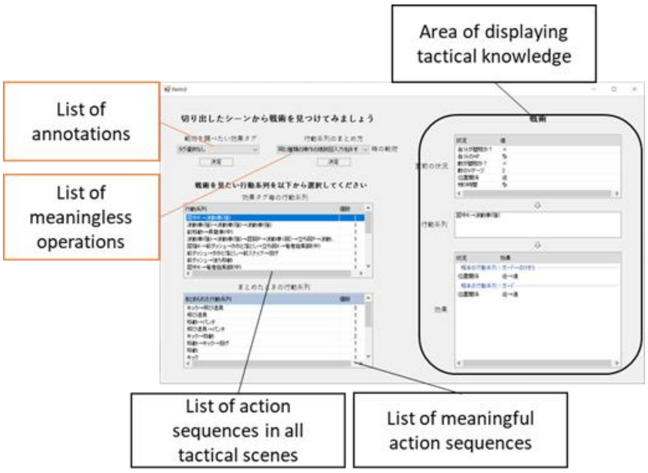

Area of displaying tactical knowledge
List of annotations
List of meaningless operations
List of action sequences in all tactical scenes
List of meaningful action sequences

Figure 5. Interface for Extracting Tactical Knowledge.

6. Evaluation Experiment

6.1 Outline

We experimentally evaluated the effectiveness of our tactical knowledge acquisition support system that obtains tactical knowledge. We prepared three play videos from Street Fighter II where the same expert used the same characters. Our research participants were 13 beginners to the Street Fighter series.

In the research, the participants watched the play video, grasped the tactical knowledge exploited by the expert, and wrote it down (Step 1). Next they used the interface to extract those scenes that they felt were good and described any tactical knowledge that they learned (Step 2). After that, they used the interface to extract tactical knowledge by browsing through the situations, actions, and effects from the cut-out scenes and added a description of new tactical knowledge that they learned (Step 3). Finally, they answered questions (Step 4).

We evaluated our system's effectiveness in terms of the comprehension of tactical knowledge and its quality. We evaluated such example of comprehension by comparing the amount of tactical knowledge described in Steps 1, 2, and 3. Regarding the quality of the tactical knowledge, grasping every element is preferred, such as action sequences, situations, and effects. Various elements of the situation must also be recognized. Therefore, evaluations are based on the amount of tactical knowledge described by situations, effects, and the number of elements in the situation.

Table 3 shows the questions asked in Step 4. Participants selected one answer from “yes,” “for the most part, yes,” “for the most part, no,” and “no.” Question 1 asks about the interface’s effectiveness for extracting scenes, and question 2 asks about its extraction of tactical knowledge.

Table 3. *Question Items*

ID	Questions
1	Was extracting the scenes easy?
2	Did the displayed sequence of actions, situations, and effects help you grasp tactics?

6.2 Experimental Results

Table 4 shows the amount of tactical knowledge each participant learned and wrote down in Steps 1, 2, and 3. Table 5 shows the answers of questions. All participants acquired new tactical knowledge in Step 3. In addition, the total number of descriptions in Steps 2 and 3 were more than double the number of descriptions in Step 1 for ten participants. The correlation coefficient was 0.422 between the number of cut-out scenes and the descriptions in Steps 2 and 3. A positive correlation was found between the number of cut-out scenes and the amount of described tactical knowledge. These results indicate that our system was effective for acquiring new tactical knowledge.

For the effects of individual interfaces, some participants noted more descriptions in Step 3 than in Step 2; some noted more in Step 2. In the answers to questions 1 and 2 in Table 5, almost all participants chose “yes” or “for the most part, yes.” Based on these results, both interfaces support the acquisition of new tactical knowledge.

Table 4. *Amount of New Tactical Knowledge*

Participant	A	B	C	D	E	F	G	H	I	J	K	L	M
Step 1	8	9	11	6	4	3	4	5	3	4	7	3	8
Step 2	4	5	0	6	0	0	3	3	4	2	1	2	2
Step 3	4	5	6	2	5	3	1	4	3	2	5	2	3

Table 5. *Answers of Questions (Number of People)*

ID	Yes	For the most part, yes	For the most part, no	No
1	5	7	1	0
2	8	5	0	0

We evaluated the quality changes of the tactical knowledge after using the system. Table 6 shows the number of elements described in the situation and the effect elements. Participants C, E, F, G, and K described both the situation and the effect after using the system, and nine of them increased the number of elements to focus on. Participant H just wrote down one action: “making a combination moves with a small movement and then hit with a big movement” before using the system as one piece tactical knowledge. After using this system, he described the situation and the effect: “when the opponent was not against the wall, push her to the wall to skillfully cause large damage.” On the other hand, participant I who did not see any change after using the system said, “To learn the tactical knowledge with the system was slightly difficult because this was the first time that I watched this game.” These results suggest that this system more effectively improved the quality of tactical knowledge for users who already had a rudimentary knowledge of the game.

Table 6. *Number of Elements of Situation and Effect Described in Tactic Knowledge*

Participant	A	B	C	D	E	F	G	H	I	J	K	L	M
a) Situation													

Step 1	3	1	2	1	0	1	2	2	2	4	3	0	2
Steps 2 and 3	4	1	2	1	2	3	2	3	3	4	3	0	2
b) Effect													
Step 1	1	0	0	0	1	0	0	1	0	2	0	0	2
Steps 2 and 3	2	0	1	0	3	1	1	4	0	3	1	0	2

7. Conclusions

We constructed a system that extracts the elements of tactical knowledge from an expert's play video scenes selected by users to support the acquisition of tactical knowledge. Based on our experimental results, using the system increased the amount and improved the quality of the tactical knowledge understood by our participants. This result suggests that extracting scenes and deriving common points from them were effective for acquiring new tactical knowledge.

This system extracts situations and effects that are common to the all tactical scenes of identical annotation. Tactical knowledge can be applied to various situations. If such knowledge does not appear as common features in the extracted play videos, this system cannot extract it. Although such features might not appear in all of the tactical scenes, they do appear in most of them. To extract the common features that appeared in many tactical scenes, a function must be created that extracts the elements in the situations and the effects seen in more than a certain number of tactical scenes.

In addition, the current system regards the common elements of situations and effects as meaningful aspects. However, not all of them are relevant to tactical knowledge. We need to investigate the important elements by interviewing experts and analyzing the relationships between each element and the game flow to create a function that only extracts meaningful elements as tactical knowledge.

References

Jenny, S. E., Manning, R. D., Keiper, M. C., & Olrich, T. W. (2017). Virtual (ly) athletes: where eSports fit within the definition of "Sport." Quest, 69(1), 1-18.

Baker, J., Cote, J., & Abernethy, B. (2003). Sport-specific practice and the development of expert decision-making in team ball sports. Journal of applied sport psychology, 15(1), 12-25.

Zhai, G., Fox, G. C., Pierce, M., Wu, W., & Bulut, H. (2005). eSports: collaborative and synchronous video annotation system in grid computing environment. In Seventh IEEE International Symposium on Multimedia.

Takahashi, K., Inomo, H., Shiraki, W., Isouchi, C., & Takahashi, M. (2017). Experience-based training in earthquake evacuation for school teachers. Journal of Disaster Research, 12(4), 782-791.

Saito, Y., Kimura, M., & Ishizaki, S., (2015). Real-time prediction to support decision-making in soccer. 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 218-225.

Wu, Y., Lan, J., Shu, X., Ji, C., Zhao, K., Wang, J., & Zhang, H. (2017). iTTVis: Interactive Visualization of Table Tennis Data. IEEE transactions on visualization and computer graphics, 24(1), 709-718.

Vales-Alonso, J., Chaves-Diéguez, D., López-Matencio, P., Alcaraz, J. J., Parrado-García, F. J., & González-Castaño, F. J. (2015). SAETA: A smart coaching assistant for professional volleyball training. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 45(8), 1138-1150.

Janusz, A., Slezak, D., Stawicki, S., & Stencel, K. (2018). SENSEI: an intelligent advisory system for the esport community and casual players. Proc. of 2018 IEEE/WIC/ACM International Conference on Web Intelligence, 754-757.

Nakagawa, A. (1984). Some basic concepts for the study on situational judgement in ball games. Japan Journal of Physical Education, Health and Sport Sciences, 28(4), pp. 287-297. [in Japanese]

Robot with Embodied Interactive Modes as a Companion Actor in Journey of Digital Situational Learning Environment and its Effect on Students' Learning Performance

Vando Gusti Al HAKIM[a], Su-Hang YANG[b], Jen-Hang WANG[c*], Chiu-Chen YEN[a], Lung YEH[d] & Gwo-Dong CHEN[a]

[a]*Department of Computer Science and Information Engineering, National Central University, Taiwan*
[b]*Department of Hospitality Management, Chien Hsin University of Science and Technology, Taiwan*
[c]*Research Center for Science and Technology for Learning, National Central University, Taiwan*
[d]*Da-Luen Junior High School, Taiwan*
*harry@cl.ncu.edu.tw

Abstract: When learning in a drama for situational learning, students need to solve the simulated problems encountered in the real world and may need a companion to guide and help them finish the drama journey. Due to the engagement, real-time feedback, and curiosity that a robot can bring, it can become an interesting actor companion to make drama performance more attractive. Additionally, physical interaction offered by the robot can make students learn through embodied interactions according to the situation and time. Therefore, this study proposes a learning approach, where student actors and the robot are immersed in digital drama scenarios, play drama, and interact with robots, scenarios, and virtual objects inside the digital situational learning environment. To evaluate the effectiveness of the proposed approach, a quasi-experiment was conducted in an English as a second language course for junior high school students. Three classes of students were assigned with different experimental instructions of learning with a robot actor in a situational learning environment. The experiment results showed that a tangible robot with more interactions significantly impacted students' learning performance. Questionnaire results revealed that students' learning motivation and engagement were improved when a robot actor with embodied interactive modes, including context-related oral interaction, touch interaction, and gesture interaction, was employed inside the digital situational learning environment.

Keywords: Digital situational learning environment, drama-based learning, situated learning, social robot, human-robot interaction, embodied interaction

1. Introduction

According to the situated cognition theory, learning should be contextualized and should not neglect the application of acquired knowledge in daily life (Brown, Collins, & Duguid, 1989). The critical characteristic of situated learning is to provide an authentic environment (Herrington & Oliver, 1995). Some previous studies demonstrated that students could effectively connect the action and knowledge of real situations when the environment of learning is similar to the actual world (Fadeeva *et al.*, 2010). As reality technologies (such as a 2D/3D scenario, Virtual Reality, and Mixed Reality) continue to develop, teachers or publishers can more easily integrate authentic scenes and scenarios pertinent to learning activities into a classroom.

Rousseau (1817) presents *learning by dramatic doing*, in which employing drama in the learning approach was introduced. Using drama in learning with a realistic environment and role-playing allows students to experience and implement their knowledge acquired from drama activities situationally. Several studies combined digital reality technology to build a digital situational learning environment (DSLE) for students to learn through drama performing and script designing (Cai *et al.*, 2020; Liu *et al.*, 2017; Wang *et al.*, 2020). It can assist teachers to build various digital authentic scenarios in accordance with textbooks to enable students to perform situational learning inside the classroom.

However, when performing situational learning, the student actors may encounter problems and may not be confident enough. Student actors might not reflect and correct their acts immediately because no one would notice their mistakes, leading to poor performance of drama learning. Consequently, they might not learn the correct knowledge or learning materials from drama scripts. Additionally, in order to engage the audiences during drama performing, talented actors are required to show apparent affect and express situations effectively (Murphy *et al.,* 2011). However, students are generally not professional actors. The actors' engagement in drama performing and audiences' interest in watching may be reduced if drama activities do not engage in situational interactions. Thus, they might need a companion or a mentor to guide, evaluate, and help them finish the drama journey.

To address the aforementioned issues, the robots can be added and designed as an actor and companion in a DSLE. One of the reasons why the robots are needed for DSLE is the curiosity (Gordon, Breazeal, & Engel, 2015) and the ability to enhance concentration and learning interest (Al Hakim *et al., 2020*). Referring to Barab *et al.* (2010) transformational play and Vogler (2007) archetypes in a drama journey, the robots can play the role of the "messenger" to tell the student actors what and how they should do in a stage of the journey. Additionally, it can also role-play as a "threshold guardian" to evaluate and correct the mistakes of student actors during drama performance, so the student actors need to focus on learning materials embedded in the drama stage. Students might develop an increased level of learning motivation through the activity and be forced to study in advance to do well in the drama performance. Promoting students' learning motivation has been recognized as an essential issue since robust learning motivation can improve learning performance (Murphy & Alexander, 2000).

Furthermore, robots' tangible and physical interaction can make student actors learn through embodied interactions, such as physical touch and oral interaction (Barnes *et al.*, 2020; Saerbeck *et al.,* 2010), and also movement or gesture (Kanda *et al.,* 2004). By utilizing related techniques, users can easily develop robot programs and integrate AI cognitive services (*e.g.*, Google Dialogflow, Microsoft Azure) into the robot and DSLE. Therefore, the learning system may have multimodal interactions and rich sensory information. The students might be highly engaged if the learning context is in the interactive mode with rich sensory information (Barnes *et al.,* 2020; Chin, Hong, & Chen, 2014).

In this study, a learning approach is proposed by adopting a robot with embodied interactive modes as a companion actor and performing situational learning with student actors in the virtual world provided by a designed DSLE. To evaluate the effectiveness of the learning approach, a situational learning activity of an English as a second language course for junior high school students was conducted to compare the learning outcomes, motivation, and engagement of the students who learned using the DSLE with robot embodied interaction approach, the conventional DSLE approach, and the conventional robot instruction approach. This study proposed the following hypothesis: DSLE with robot embodied interactive modes can enhance learning outcomes, motivation, and engagement.

2. Literature Review

2.1 Reality-Assisted Learning Technology

Reality technology is nothing new nowadays, but related technologies have developed on several research for years. For instance, Microsoft released *Kinect for Windows v2* in July 2014 and provided a new version of its SDK for the platform (https://developer.microsoft.com/en-us/windows/kinect/). With motion detection technology, a virtual world with embodied immersion can be easily and quickly constructed. By standing in front of Kinect v2, users' body positions can be retrieved through skeleton coordinates, which enables their full-body elements immersed into the digital scenario and wearing digital props, costumes, or objects (Liu *et al.,* 2017). Furthermore, users can watch their images in real-time displayed on the screen and reflect their performance in the virtual world (Wu *et al.*, 2015).

Several studies extended the reality-assisted learning technology with embodied interaction to promote learning performance and enhance students' experience in situational learning inside the classroom. For instance, the study of Wang *et al.*, (2020) employed body movements and considered the impacts of social cognition inside the classroom. The study results revealed that the students could simultaneously get assessments and corrections in real time, which significantly promotes students' learning performance. Besides, the learning gains and enhancement of students' motivation can also be cultivated by using mechanisms of real-time spoken language evaluation (Cai *et al.,* 2020). However, human-robot interactions that allow students to have embodied interactions through tangible and physical interaction were not addressed in the aforementioned studies.

2.2 Learning Environment with Robots

Employing robots into the classroom environment enables robots to perform as learning partners (Kanda *et al.,* 2004) or teacher assistants (Chang, Lee, Chao, Wang, & Chen, 2010; Chin, Hong, & Chen, 2014), which mostly emphasizes having a positive impact on learning. Physical robot interaction can facilitate students to get instant feedback (Chang, Lee, Wang, & Chen, 2010). Barnes *et al.,* (2020) proposed a child-robot theater framework by employing several robots to engage elementary students. In the process of theater preparation, teacher and students or developers may need to set up the props, scenery, costumes, and the narrative in order to play theater in the most authentic way (Romero-Hernandez *et al.,* 2018). Mixed-reality can be an interesting option to build props and stages (Bravo Sánchez *et al.,* 2017), which are difficult to be obtained in reality according to the learning content (Liu *et al.,* 2017).

Regarding related mixed-reality technology, importing robots into mixed-reality environments can construct a hybrid physical-digital user interface, which can be beneficial for learning (Antle & Wise, 2013). For instance, the prior study of Chang, Lee, Wang, and Chen (2010) presented a RoboStage system by importing robots into authentic mixed-reality. They claimed that robots could make the learning activities more engaging and enjoyable, which consequently affected students' motivation. It also concluded that students have a stronger preference to interact with a robot than virtual interactions. Nevertheless, such robot interaction in mixed-reality technology aforesaid was constrained to the physical environment only. In other words, the robots were not involved inside the digital scenario tasks.

Cheng, Wang, and Chen (2019) presented a design framework to guide researchers or teachers in developing an immersive language environment by utilizing tangible robots and IoT-based toys. Each presented important key points of design principles and guidelines can be as references to build an immersive language environment. Such an environment allows learners to play with the IoT toys while being accompanied by robots that provide them with parent-like linguistic feedback but done in the target language. However, researchers did not address a digital situational scenario to provide learners an authentic context or virtual world quickly and easily inside the classroom.

3. System Design and Implementation

DSLE with robot embodied interaction is constructed to assist teacher and students learning inside the classroom. It enables students and a robot together to role-play in various situations and provides a real-time evaluation under the situation. The teacher can choose interaction modes and virtual objects to quickly design the digital scenarios based on textbooks' content and contexts and put learning materials inside the scenario. Learning materials (*e.g.,* keywords, actions, and postures) are applied in a scenario, and students must learn them so they can confront challenges within the journey stage.

3.1 System Architecture

In this study, the display interface of DSLE is constructed by C#, Windows Presentation Foundation (WPF), and .NET Framework 4. Xbox Game Bar is used for video recording during drama performances. In addition, Zenbo SDK is used to develop robot capabilities (*e.g.,* touch sensing, emotions, and mobility). Besides, to enhance the capabilities of robots, Microsoft Azure Cognitive Speech Services API and Google DialogFlow API were involved in managing the speech recognition. Also, Microsoft Kinect V2.0 and its SDK were adopted to extend robots' eyes for capture and recognize actors' body skeletons. The main control application and robot program are developed on the Android platform. All of the system components communicate through a socket connection.

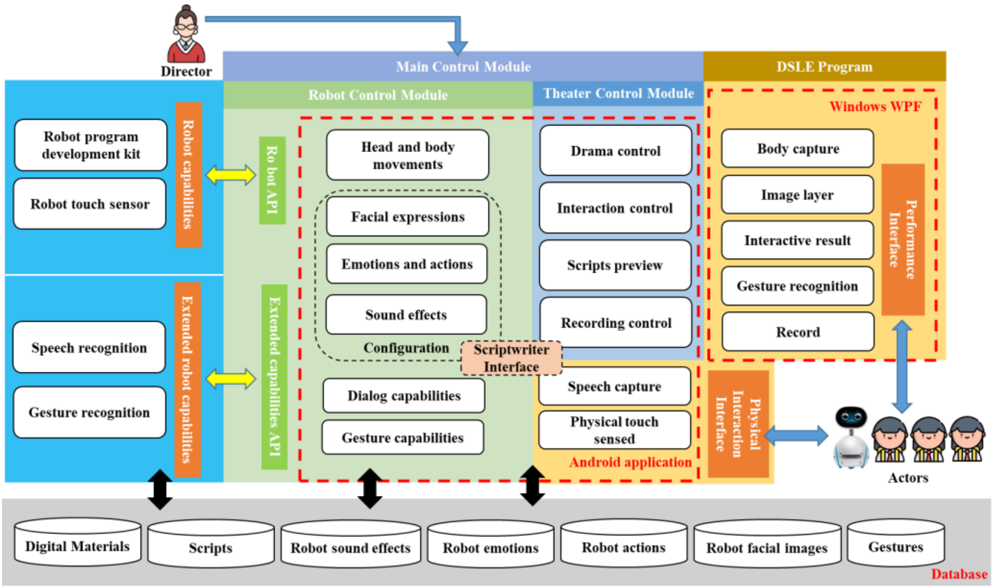


Figure 1. System Architecture.

Figure 1 depicts the system architecture of the study, which majorly consists of the main control module installed on a tablet, and a DSLE program installed on a personal computer with an NVIDIA graphics card. During drama performance, the teacher can situationally control the transition of digital scenes and the robot through the tablet. The main control module controls "Theater" (*e.g.*, scenes shift, interactions, scripts preview, display subtitles, and recording) and "Robot" (*e.g.*, head and body movements). Meanwhile, the DSLE program installed on the computer runs with Kinect v2 to capture actors' image and body skeleton information, and then construct a virtual world by combining scenarios with digital materials.

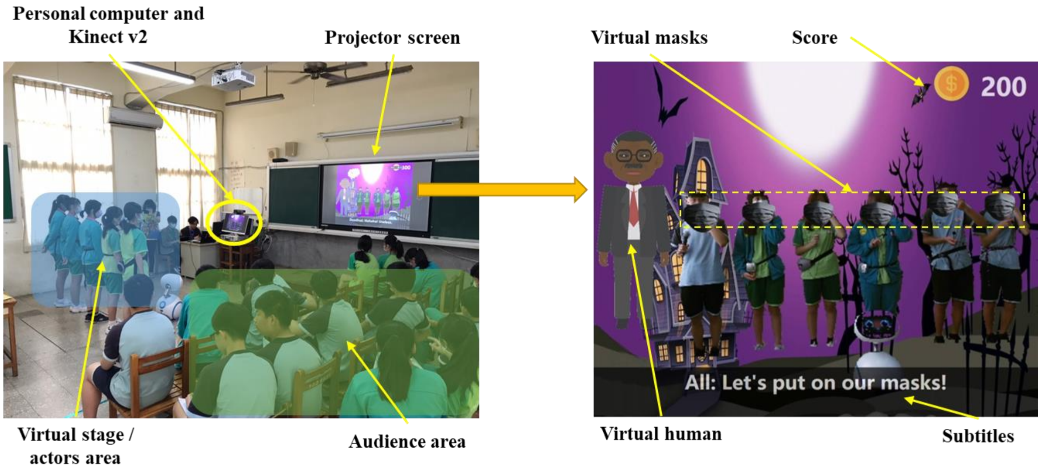
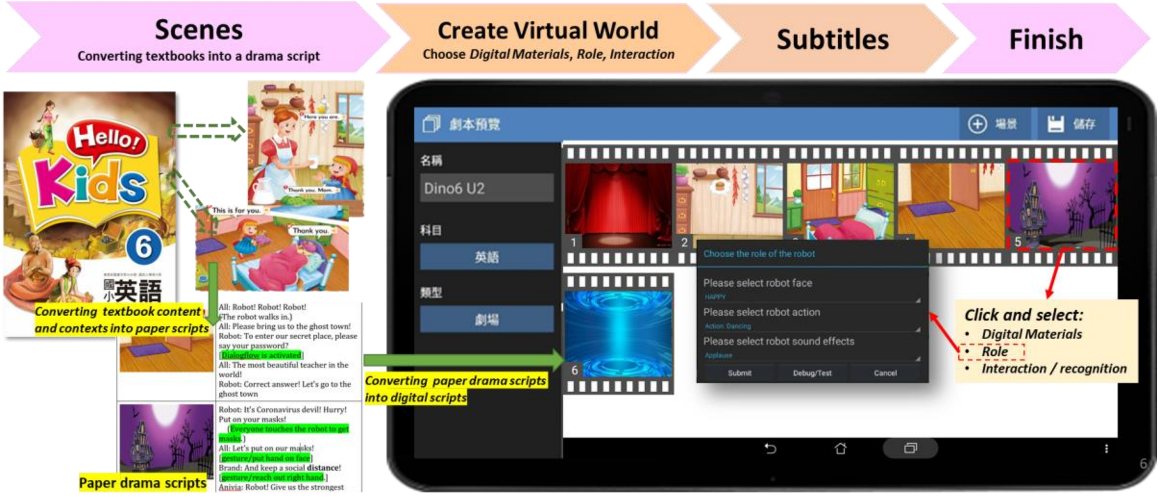

Personal computer and Kinect v2
Projector screen
Virtual masks
Score
200
All: Let's put on our masks!
Virtual stage / actors area
Audience area
Virtual human
Subtitles

Figure 2. Composition of DSLE with Robot Embodied Interactive Modes.

Figure 2 displays the composition of the DSLE with robot embodied interactive modes. In this study, the classroom is separated into a virtual stage area and an audience area. By standing in front of Kinect v2, the robot and a group of student actors are immersed into the digital drama scenarios, play drama, and have interactions with a robot, scenarios, and virtual objects inside it. Student actors can watch their own performance and get instant feedback from the robot and virtual world through a computer screen, so learning activities become interactive. The computer screen in the virtual stage area is projected to the screen in the audience area, so the rest of the students can watch and learn actor performances through live broadcast. The learning system through the computer screen will display the virtual world and the task or mission based on script design to encourage actors to move, act, and speak narratively.

In order to make the robots able to do the same actions as student actors do, system communication of DSLE is extended by adding a database host system with robot materials (*i.e.*, sound effects, emotions, actions, facial images). By doing so, the robot can connect with the main control module on the tablet and DSLE program through socket communication. Consequently, before drama performing, the teacher can use scriptwriter interface (Figure 3) to write plots and set up the digital scenarios (*e.g.*, background, foreground, masks, and props) and programs the robots' role (*i.e.*, actions, facial emotions, and sound effects) on each stage of the journey based on the designed script.



All: Robot! Robot! Robot!
(The robot walks in.)
All: Please bring us to the ghost town!
Robot: To enter our secret place, please say your password?
[Dialogflow is activated]
All: The most beautiful teacher in the world!
Robot: Correct answer! Let's go to the ghost town

Robot: It's Coronavirus devil! Hurry! Put on your masks!
[Everyone touches the robot to get masks.]
All: Let's put on our masks!
[gesture/put hand on face]
Brand: And keep a social distance!
[gesture/reach out right hand.]
Anivia: Robot! Give us the strongest

Choose the role of the robot
Please select robot face: HAPPY
Please select robot action: Action: Dancing
Please select robot sound effects: Applause
Submit | Debug/Test | Cancel

Figure 3. Scriptwriter Interface of the Main Control Module.

3.2 *Teaching Procedure of DSLE with Robot Actor*

At first, the teacher taught the students the content of the English textbook and explained the meaning of each sentence. Next, a description of the DSLE with the robot actor was presented, including the explanation of the drama script and demonstration. The teacher then provided students with instructions regarding the drama script, assisted them to practice the script (*e.g.*, moves or acts), and helped them to learn pronunciation. Afterwards, students are divided in small groups (5 to 7 each) to rehearsal using the system in turn. In addition, they might practice performing with a paper drama script after class. Following the rehearsals, actor groups performed a formal drama show using the system. The other classmates watching the show as an audience could reflect and learn from the actors. As each actor group had finished its performance, the teacher discussed the performance and let other classmates give advice to the actors. Finally, the drama performance videos were uploaded to the database for evaluation and reflection after class.

3.3 *Interaction Design*

Utilizing the physical interface of the robot, cognitive services, and DSLE, several modes of embodied interaction can be created. The interaction modes were integrated into a particular stage of the journey based on the drama script. It consists of context-related oral interaction with the robot, robot physical touch interaction, and gesture interaction. The teacher can insert some important learning materials into the learning drama script as a challenge from the robot that acts as a threshold guardian for student actors to get information, virtual objects, or pass a gate to the next journey stage. The robot actor will always give a command and instruction until actors can accomplish the challenge. Real-time mirrored feedback from the virtual world, and instant feedback from the robot are employed, so that the actors can get a real-time evaluation and remedial action during drama performance. The evaluation and real-time assessment will enforce students to learn the drama script in advance so that they can perform well in the learning drama with the robot actor.

The system will recognize actors' interactions and give real-time evaluation. When the system verifies that the actor has properly done his task based on the learning materials, the robot will show positive feedback (*e.g.*, happiness, smile, and joy), and the virtual world will also give a score to them and attach a virtual object or flow to next stage of the journey. On the other hand, if the actor makes a mistake during the recognition period, the robot will show negative feedback (*e.g.*, sadness, cry, and anger) and encourage them to try again. In such a situation, the virtual world will decrease the score and attach a burned face mask effect to actors. Example of the recognition process is shown in Figure 4.

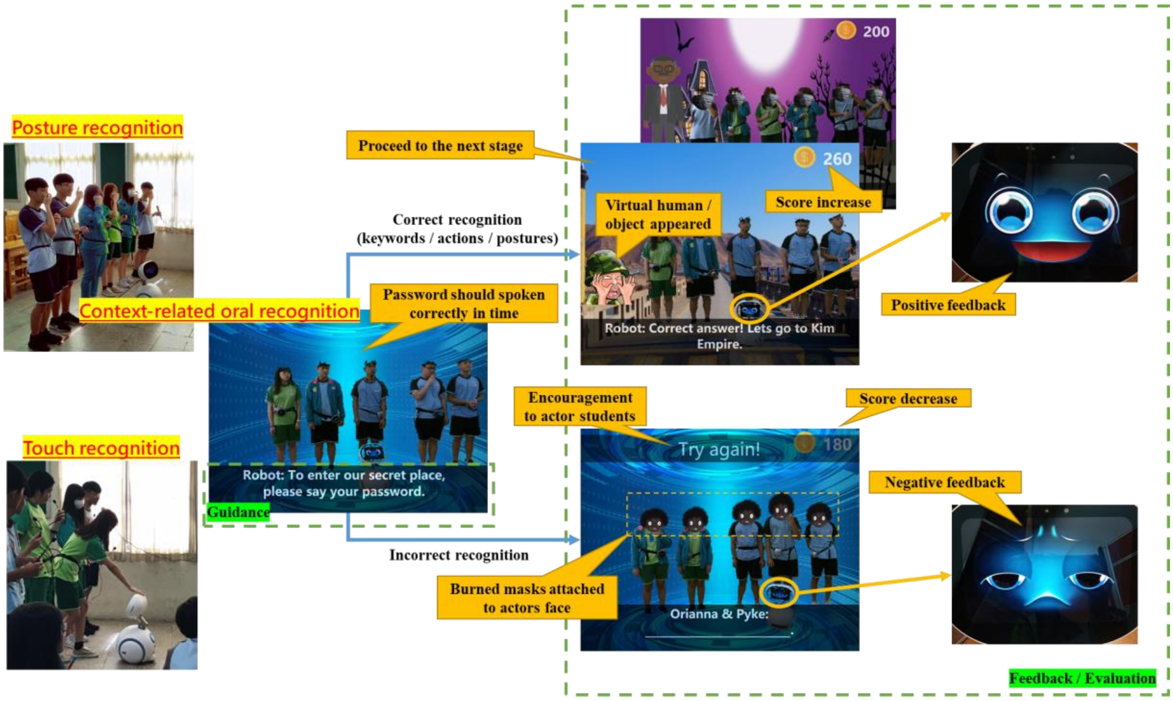


Figure 4. Example of the Recognition Process.

4. Experiment

4.1 Participants & Learning Materials

In a junior high school English as a second language course in Taiwan, this experiment was conducted. There were three classes randomly selected to become the experimental group A, experimental group B, and control group, respectively. All of the students in groups were grade 9 (around 15 years old) and taught by the same teacher. Experimental group A containing 39 students (25 males, 14 females) used the DSLE with robot embodied interactive modes, in which robot and student actors were immersed into a digital scenario and have physical robot touch interaction, gesture interaction, and context-related oral interaction. The experimental group B consisting of 31 students (18 males, 13 females) used the conventional DSLE, in which robot and student actors were only immersed into a digital scenario and had no recognition. In other words, in experimental group B, students only did drama performance with less interactions of robot actor (*i.e.*, the robot actor only spoke the subtitles based on context-related same as student actors did). The other class with 31 students (20 males, 11 females) was selected as the control group learned with conventional robot instruction approach, in which students used the display screen as social robots had in general and were not immersed into a digital scenario. Details of the experimental group A, experimental group B, and control group, with decreasing levels of interaction and immersion, respectively, are shown in Table 1.

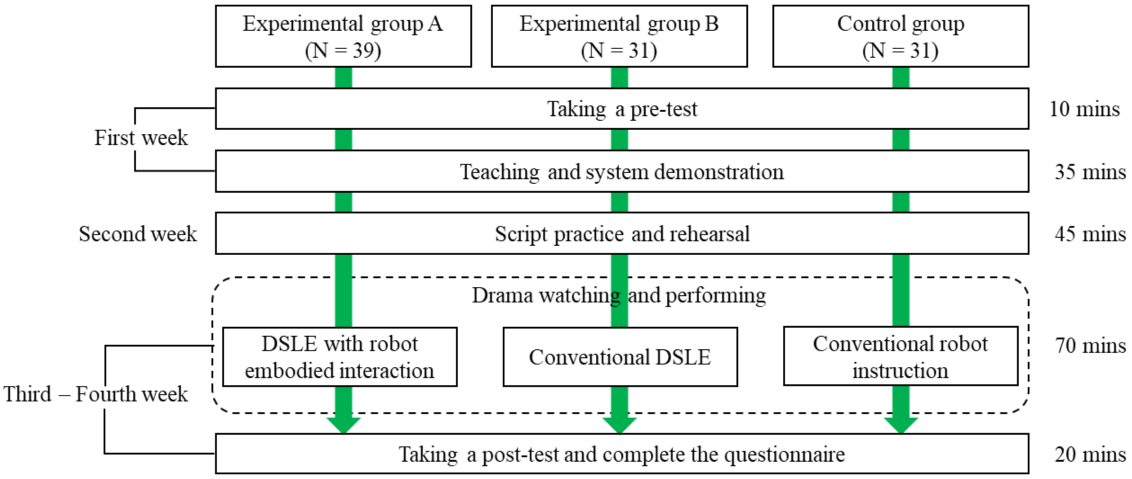
Table 1. *How Students and Robots were Interacting with Each Other in Each Group*

Group	Immerse	Touch recognition	Gesture recognition	Context-related oral recognition
Experimental group A	✓	✓	✓	✓
Experimental group B	✓	✗	✗	✗
Control group	✗	✗	✗	✗

This study compiled the materials taken from the English course textbook into a drama script. The drama script was developed by a school teacher with considerable experience in teaching English (more than 10 years). Additionally, the drama script screenwriting was based on the journey structure proposed by Vogler (2007) and Barab *et al.* (2010) to make the developed drama script convey the narrative structures and character development. All three groups used the same drama script.

4.2 Evaluation

This study was conducted for one month. A pre-test for total score ranged from 0 to 100 points developed by the teacher was administrated at the beginning of the experiment to evaluate the prior knowledge level of students. In order to keep the test results consistent, a post-test similar to but different from the pre-test was conducted at the end to examine students' learning outcomes. Based on the aforementioned teaching procedure and experiment subjects, the experimental design was developed, as shown in Figure 5.



Experimental group A (N = 39)
Experimental group B (N = 31)
Control group (N = 31)
First week
Taking a pre-test
10 mins
Teaching and system demonstration
35 mins
Second week
Script practice and rehearsal
45 mins
Drama watching and performing
DSLE with robot embodied interaction
Conventional DSLE
Conventional robot instruction
70 mins
Third – Fourth week
Taking a post-test and complete the questionnaire
20 mins

Figure 5. Experimental Design.

Based on the test scores obtained in this study, a single factor covariate analysis (ANCOVA) was used to compare the learning performance between the three groups. Additionally, a questionnaire with a five-point *Likert* scale was employed to examine students' learning motivation and engagement. The questionnaire that was used to evaluate the students learning motivation after the treatment was developed based on Hwang, Yang, and Wang (2013). It consists of 7 items questions (*e.g.*, "I will practice the script seriously when participating in the scenario of the script with the robot together in a drama journey") with the *Cronbach alpha* value proposed by the original study was 0.79 (N = 56). While for evaluating the learning engagement, the emotional engagement measure of Jamaludin and Osman (2014) was adopted. It consists of 5 items (*e.g.*, "Interacting with the robot in a drama journey can deepen my impression of the learning content") and has been reported a reliability coefficient (*Cronbach alpha*) of 0.955 (N = 24). Since the Taiwanese students had a low proficiency level, all items were translated to Mandarin.

5. Results and Discussion

5.1 Learning Performance Analysis

The ANOVA test was conducted before ANCOVA to verify whether the groups had similar prior knowledge. ANOVA test result ($F = .367$ with $p > .05$) indicated that there was no significant difference in the pre-test, suggesting that all three groups possessed the same prior knowledge before the experiment. Then, the ANCOVA analysis was performed using the pre-test scores as covariance and the post-test scores as the dependent variable.

The result of ANCOVA shows that the adjusted mean of the experimental group A, experimental group B, and the control group were 79.54, 64.45, and 72.51, respectively. Once the pre-test impact was removed, a significant difference emerged between the three groups with $F = 11.58$ ($p < .05$), demonstrating that significant differences in post-test scores were found between the three groups. By applying pairwise comparisons, it was found that experimental group A led both experimental group B and control group, demonstrating that the group of DSLE with robot embodied interactive modes had significantly higher learning outcomes than the conventional DSLE modes and the conventional robot instruction modes.

Students can learn better when the embodied interactive modes (*i.e.*, context-related oral interaction, gesture interaction, and physical touch interaction) are added into the robot actor and DSLE. Due to the interactive and instant feedback assessment provided by the virtual world and a robot with embodied interactions, students learn the drama script seriously to perform well in the drama journey, so they outperform those in the other group in terms of learning performance. As stated by Kerawalla *et al.* (2006), real-time evaluation and corrections during learning activities are essential. Thus, we concluded that more recognition and interaction in the situational learning environment could encourage students to study in advance.

Unexpectedly, pairwise comparison results also demonstrate that the conventional robot instruction group outperformed the conventional DSLE group in terms of learning outcomes. Possibly, this finding was caused by the students in the conventional DSLE group not participating in the direct interaction with the robot actor, but instead interacting with the virtual robot actor through the computer screen. Even though the conventional robot instruction group students were not immersed in the digital scenarios, they directly interact with the tangible robot actor. Based on previously published studies (Belpaeme *et al.*, 2018; Chang, Lee, Wang, & Chen, 2010; Leyzberg *et al.*, 2012), physical robots are more effective at fostering learning gains than virtual robots. In line with that, Ahtinen & Kaipainen (2020) stated that having robots physically present in a classroom contributes immensely to classroom interaction and learning.

5.2 Questionnaire Analysis

Questionnaire responses were analyzed with ANOVA to compare the students' learning motivation and engagement after received the different approaches. The questionnaire contained 101 responses, and 100 of them were considered valid copies. The ANOVA result of students' learning motivation shows that experimental group A scored 4.05, experimental group B scored 3.33, and the control group scored 2.97. Moreover, significant differences $F = 23.92$ ($p < .05$) were found among the three groups regarding learning motivation. Pairwise comparison showed that experimental group A was superior to experimental group B and the control group, implying that the learning motivation is higher for the DSLE with robot embodied interactive modes compared to the other two groups. It is suggested that robot-based learning systems that combine physically embodied robots with instructional tools or attractive multimedia objects inside the classroom can provoke a strong motivation to learn (Chang, Lee, Chao, Wang, & Chen, 2010; Chin, Hong, & Chen, 2014).

Moreover, it was also found that both of the DSLE with embodied and non-embodied interactive modes of robot actor could significantly improve the students' learning motivation in comparison with the conventional robot instruction mode. As claimed by previous studies (Liu *et al.*, 2017; Wu *et al.*, 2015), we believe that the DSLE can help to increase the novelty effect and interaction in situational learning so that the students can keep their learning motivation and fondness.

In terms of learning engagement, several previous studies (Fitter & Kuchenbecker, 2019; Verner *et al.*, 2016; Saerbeck *et al.*, 2010) claimed that robots are suggested to make interactions more fun and engaging. A mean score of 4.05 was obtained from the experimental data analysis in experimental group A, while 2.77 was achieved in experimental group B, and 2.90 was obtained in the control group. Based on the obtained F-value ($F = 22.61$ with $p < .05$), significant differences existed between the learning engagement of the three groups.

Pairwise comparisons revealed that experimental group A outperformed experimental group B and control group. However, no significant difference was found between the experimental group B and the control group. The ANOVA result indicates that the DSLE with robot embodied interaction group outperformed the other two groups in terms of learning engagement, while the learning engagement of the conventional DSLE and the conventional robot instruction group did not have a significant difference. This may happen due to the students in DSLE with robot embodied interaction group have to perform robot actor' challenges in order get information, virtual objects, or pass through a gate to the next journey stage, which will be fun and engaging for them. We conclude that the immediate feedback support that was imposed on the learning system led to increased engagement in learning (Yeh, Cheng, Chen, Liao, & Chan, 2019).

6. Conclusion

The work proposed a situational learning environment with a robot as a companion actor together with students role-playing in the journey inside the digital scenario. The robot actor was designed and programmed to do the same acts as human actors do in situational learning activities. Both robot and student actors are immersed into the digital scenarios, play drama, and have interactions with robot, scenarios, and virtual objects inside DSLE. The major role of the robot in the script is to play as a messenger to provide guidance, and at the same time as the threshold guardian, thus using the learning materials is necessary for students to complete a task and move to the next part of the drama journey. The robot with cloud AI capabilities and DSLE offers real-time feedback to evaluate whether students learn the knowledge.

Based on the experimental results, it is concluded that students can learn better when the embodied interactive modes (*i.e.*, context-related oral interaction, gesture interaction, and physical touch interaction) are added into the robot actor and DSLE. More interactions provided by the proposed system could enforce students to study in advance in order to perform well in the drama performance. Furthermore, our findings demonstrate that the robot embodied interactive modes can promote learning motivation and engagement of learning inside DSLE.

Acknowledgements

The author would like to thank all the teachers and students who were involved in this experiment. This study is supported in part by the Ministry of Science and Technology of the Republic of China under contract numbers MOST 109-2511-H-008-003-MY2, MOST 109-2511-H-008-004-MY3, and MOST 110-2811-H-008-505.

References

Ahtinen, A., & Kaipainen, K. (2020, April). Learning and teaching experiences with a persuasive social robot in primary school–findings and implications from a 4-month field study. In *International Conference on Persuasive Technology* (pp. 73-84). Springer, Cham.

Al Hakim, V. G., Yang, S. H., Tsai, T. H., Lo, W. H., Wang, J. H., Hsu, T. C., & Chen, G. D. (2020, July). Interactive Robot as Classroom Learning Host to Enhance Audience Participation in Digital Learning Theater. In *2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT)* (pp. 95-97). IEEE.

Antle, A. N., & Wise, A. F. (2013). Getting down to details: Using theories of cognition and learning to inform tangible user interface design. *Interacting with Computers*, *25*(1), 1-20.

Barab, S. A., Gresalfi, M., & Ingram-Goble, A. (2010). Transformational play: Using games to position person, content, and context. *Educational researcher*, *39*(7), 525-536.

Barnes, J., FakhrHosseini, S. M., Vasey, E., Park, C. H., & Jeon, M. (2020). Child-robot theater: Engaging elementary students in informal STEAM education using robots. *IEEE Pervasive Computing*, *19*(1), 22-31.

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. *Science robotics*, *3*(21).

Bravo Sánchez, F. Á., González Correal, A. M., & Guerrero, E. G. (2017). Interactive drama with robots for teaching non-technical subjects. *Journal of Human-Robot Interaction*, *6*(2), 48-69.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational researcher*, *18*(1), 32-42.

Cai, M. Y., Wang, J. Y., Chen, G. D., Wang, J. H., & Yang, S. H. (2020, July). A Digital Reality Theater with the Mechanisms of Real-Time Spoken Language Evaluation and Interactive Switching of Scenario & Virtual

Costumes: Effects on Motivation and Learning Performance. In *2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT)* (pp. 295-299). IEEE.

Chang, C. W., Lee, J. H., Chao, P. Y., Wang, C. Y., & Chen, G. D. (2010). Exploring the possibility of using humanoid robots as instructional tools for teaching a second language in primary school. *Journal of Educational Technology & Society*, *13*(2), 13-24.

Chang, C. W., Lee, J. H., Wang, C. Y., & Chen, G. D. (2010). Improving the authentic learning experience by integrating robots into the mixed-reality environment. *Computers & Education*, *55*(4), 1572-1578.

Cheng, Y. W., Wang, Y., & Chen, N. S. (2019). A framework for designing an immersive language learning environment integrated with educational robots and IoT-based toys. *Foundations and Trends in Smart Learning*, 1-4.

Chin, K. Y., Hong, Z. W., & Chen, Y. L. (2014). Impact of using an educational robot-based learning system on students' motivation in elementary education. *IEEE Transactions on learning technologies*, *7*(4), 333-345.

Fadeeva, Z., Mochizuki, Y., Brundiers, K., Wiek, A., & Redman, C. L. (2010). Real‐ world learning opportunities in sustainability: from classroom into the real world. *International Journal of Sustainability in Higher Education*.

Fitter, N. T., & Kuchenbecker, K. J. (2019). How does it feel to clap hands with a robot?. *International Journal of Social Robotics*, 1-15.

Gordon, G., Breazeal, C., & Engel, S. (2015, March). Can children catch curiosity from a social robot?. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 91-98).

Herrington, J., & Oliver, R. (1995). Critical characteristics of situated learning: Implications for the instructional design of multimedia.

Hwang, G. J., Yang, L. H., & Wang, S. Y. (2013). A concept map-embedded educational computer game for improving students' learning performance in natural science courses. *Computers & Education*, *69*, 121-130.

Jamaludin, R., & Osman, S. Z. M. (2014). The use of a flipped classroom to enhance engagement and promote active learning. *Journal of education and practice*, *5*(2), 124-131.

Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: A field trial. *Human–Computer Interaction*, *19*(1-2), 61-84.

Kerawalla, L., Luckin, R., Seljeflot, S., & Woolard, A. (2006). "Making it real": exploring the potential of augmented reality for teaching primary school science. *Virtual reality*, *10*(3-4), 163-174.

Leyzberg, D., Spaulding, S., Toneva, M., & Scassellati, B. (2012). The physical presence of a robot tutor increases cognitive learning gains. In *Proc. of the annual meeting of the cognitive science society* (Vol. 34, No. 34).

Liu, Y. T., Lin, S. C., Wu, W. Y., Chen, G. D., & Chen, W. (2017). The digital interactive learning theater in the classroom for drama-based learning. In *Proceedings of the 25th International Conference on Computers in Education* (pp. 784-789). Asia-Pacific Society for Computers in Education.

Murphy, P. K., & Alexander, P. A. (2000). A motivated exploration of motivation terminology. *Contemporary educational psychology*, *25*(1), 3-53.

Murphy, R., Shell, D., Guerin, A., Duncan, B., Fine, B., Pratt, K., & Zourntos, T. (2011). A Midsummer Night's Dream (with flying robots). *Autonomous Robots*, *30*(2), 143-156.

Romero-Hernandez, A., Riojo, M. G., Díaz-Faes-Perez, C., & Manero-Iglesias, B. (2018). The Courtesy of Spain: Theater for the New Generations. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, *13*(3), 102-110.

Rousseau, J. J. (1817). *Emile* (Vol. 2). A. Belin.

Saerbeck, M., Schut, T., Bartneck, C., & Janse, M. D. (2010, April). Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1613-1622).

Verner, I. M., Polishuk, A., & Krayner, N. (2016). Science class with RoboThespian: using a robot teacher to make science fun and engage students. *IEEE Robotics & Automation Magazine*, *23*(2), 74-80.

Vogler, C. (2007). *The writer's journey*. Studio City, CA: Michael Wiese Productions.

Wang, J. H., Chen, Y. H., Yu, S. Y., Huang, Y. L., & Chen, G. D. (2020, July). Digital Learning Theater with Automatic Instant Assessment of Body Language and Oral Language Learning. In *2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT)* (pp. 218-222). IEEE.

Wu, W. Y., Luo, Y. F., Huang, D. Y., Huang, C. W., Peng, Y. I., & Chen, G. D. (2015). A Self-Observable Learning Cinema in the Classroom. In *The 23rd International Conference on Computers in Education* (pp. 257-262). Asia-Pacific Society for Computers in Education.

Yeh, C. Y., Cheng, H. N., Chen, Z. H., Liao, C. C., & Chan, T. W. (2019). Enhancing achievement and interest in mathematics learning through Math-Island. *Research and Practice in Technology Enhanced Learning*, *14*(1), 1-19.

A RECIPE for Teaching the Sustainable Development Goals

Ma. Mercedes T. RODRIGO*, Walfrido David A. DIY, Abigail Marie T. FAVIS, Francesco U. AMANTE, Janina Carla M. CASTRO, Ingrid Yvonne HERRAS, Juan Carlo F. MALLARI, Kevin Arnel MORA, Johanna Marion R. TORRES & Ma. Assunta C. CUYEGKENG
Ateneo de Manila University, Philippines
*mrodrigo@ateneo.edu

Abstract: Every individual is responsible for sustainable development and therefore awareness of sustainability should begin at a young age, when mindsets and habits are formed. One form which sustainability education takes is computer-based games. We describe how we used Nicholson's Meaningful Gamification framework to design *For People and Planet: An SDG Adventure*, a narrative-based adventure game that teaches middle school children about the United Nations Sustainable Development Goals. In this paper, we describe the framework's six elements--play, exposition, choice, information, engagement, and reflection--and show how our design choices align with these elements.

Keywords: Game design, meaningful gamification, narrative game, sustainable development goals

1. Context

As early as 1972, the Club of Rome's *Limits to Growth* (Meadows et al. 1972) already called for a "world system… that is sustainable." Eventually, there were attempts to reconcile economic development with environmental integrity, leading to the concept of sustainable development (Purvis, Mao, & Robinson, 2019). However, convincing individuals to commit to and act for sustainability is not easy because this often challenges them to sacrifice some convenience and comfort. Awareness about sustainability should start at a young age, when mindsets and habits are being developed. It is also important that individuals view sustainability within their own context because ultimately, sustainability solutions begin with the self.

Most of the sustainability aspirations were articulated in the Millennium Development Goals (MDGs) for the period 2000-2015 (UN, 2015). The MDGs became the starting point for the 17 UN Sustainable Development Goals (SDGs) developed in 2015 "to address urgent global challenges over the next 15 years" (UN, 2016). Based on these goals, the 2030 Agenda for Sustainable Development was crafted as a roadmap involving different players and sectors. After four years of implementation, however, there are gaps in the awareness of these goals and how every sector and every individual could actually contribute to the achievement of the SDGs. While governments are responsible for developing national strategies to fulfill this commitment, all sectors of society must recognize that they too have a role to play in ensuring that the SDGs are achieved.

To help bridge the gap between the general public and the SDGs, different groups have created materials such as the SDG Academy, affiliated with the Sustainable Development Solutions Network (SDSN), which has developed publicly accessible learning modules, generally designed for the adult learner. The World's Largest Lesson, in partnership with the United Nations Children's Fund (UNICEF) and the United Nations Educational, Scientific and Cultural Organization (UNESCO), has developed resources that can be integrated in formal education systems - an important conduit since schools play a role not just in educating students about the SDGs but also in guiding their formation towards the practice of sustainability in their everyday lives (Filho et al., 2019, Nousheen et al., 2020).

Because the 17 SDGs can be overwhelming, the non-profit organization Business for Sustainable Development (formerly Philippine Business for the Environment) clustered them into five thematic areas: natural capital, food systems, social services, livable communities, and ethics and governance.

However, there is a dearth of materials that show the interconnections of systems, which is inherent to the idea of sustainability, and which can be used for Philippine learners. The purpose of this paper is to describe an attempt to fill in the gap for good sustainability learning materials that would connect, not just cognitively, but also emotionally, to a young Philippine audience. We created a narrative, contextualized, computer-based game and accompanying learning materials that illustrate these five themes. In the succeeding sections, we describe existing sustainability games and how they provided inspiration for our own effort, the theoretical basis for the design decisions that we made, and the ways in which we implemented our instructional messages. This paper does not include the game evaluation because, as of the time of its writing, the evaluation was still ongoing.

2. Existing Games for Sustainability Education

Teaching about the SDGs can be daunting because it requires adjustments in existing curricula and teaching methods (Williamo et al, 2018, Hensley, 2020). Thus, innovative approaches are needed since traditional methods have low engagement, especially with young learners (Prensky, 2006). Using games as an educational tool has been shown to be an efficient approach to explain environment and sustainability related concepts (Bevilacqua et al., 2015; Katsakiali & Mustafee, 2012).

Several games for sustainability education have been developed. The *2030 SDGs Game*, developed by Immacocollabo (Japan) in 2016 is a card game that can be played by 5 to 50 people. It is a simulation game for adult learners (practitioners in government, industry, and other institutions), and is an exercise in collaborative engagement for sustainability (Immacocollabo, 2018). In a similar vein, *The World's Future* is a simulation game also for adult learners that shows the realities and complexities of achieving sustainability given the different contexts and situations that exist (Centre for Systems Solutions - CRS, 2019).

Go Goals is an educational board game for children that introduces them to the 17 SDGs and how these affect their lives. Created by the United Nations, the gameplay is similar to *Snakes and Ladders*, and all materials (board game sheet, dice, tokens, information cards) are downloadable for free (UN, 2019).

Once Upon a Tile is a prototype mobile video game the Android, iOS, Windows Phone, PC, Mac, and Linux platforms. Developed by We Are Muesli, Pietro Polsinelli and Daniele Giardini, the objective of the game is to match resource tiles to achieve peace and sustainable development (We are Muesli, n.d.).

World Rescue, developed by Zu Digital, is a mobile video game for the Android and iOS platforms. The game is meant for a younger audience and is set in Kenya, Norway, Brazil, India, and China. In the game, the player assists non-player characters to solve issues related to displacement, health, deforestation, and pollution (Zu Digital, 2020).

Perhaps the most mainstream game for sustainability is the *Oil Springs* expansion to the popular board game "Settlers of Catan" by Klaus Teubers. In this expansion, developed by Erik Assadourian and Ty Hansen, players have to manage the advantages of having a fossil fuel resource against the potential negative impacts of its extraction and use (Catan GmBH, 2020).

The games we reviewed have several recurring themes. They contextualize the SDGs in the day-to-day and they challenge players to solve problems through proper allocation and management of available resources. The games are not tutorials disguised as games. Rather, they are activity-oriented and they persuade the players the fulfillment of the SDGs is not simply the work of individuals, not just of large institutions. To develop the game described in this paper, we borrow from the game patterns and themes used in prior games, specifically those of contextualization in the day-to-day, problem solving, and empowerment.

3. For People and Planet: An SDG Adventure

For People and Planet: An SDG Adventure is a narrative-based adventure game for the mobile phone or tablet that helps learners see the SDGs in their day-to-day lives. The adventure game genre was chosen because of its emphasis on story and exploration. Players assume the role of a middle school student in a rural community in the Philippines.

The game itself is divided into five (5) stories that can be played in any order. In each story, the player is tasked with performing errands and other everyday activities while also being shown how the community encourages sustainability and maintains their environment. Each story covers an aspect of life and ties it to one or more SDGs. In *What's For Lunch?*, the player is asked to buy food for the family's lunch. Afterwards, the player's grandmother asks for assistance with meal preparation and food waste disposal. In *Flood Fighters*, the player accompanies the grandmother and learns about disaster risk reduction and disaster risk management, both at the community and at the household levels. In *A Walk in the Park*, the player goes to a neighboring town with friends to learn about clean energy and wastewater treatment. On their way to visit the town park's bird sanctuary, the group also makes friends with another visitor who was bullied for her physical disability. In *Work, Work*, the player visits the community enterprise that the grandmother works in. Various employees of the enterprise tell the player about their work, which includes ecotourism in the nearby coastal area. In *Learning is for Everyone*, after attending a theater workshop in school, the player befriends a student with low vision. Together, they search the school for the whereabouts of their grandparents and engage in a variety of activities pertaining to the education system. Gameplay involves exploring areas and playing mini-games to fulfill quests.

Information is presented mostly through small chunks of text accompanied by colorful background images in a format similar to a visual or kinetic novel. Mini-games are designed to be a simple and fun way to break the monotony of reading character dialogue and educational text. Some mini-games also provide an alternative way to present certain lessons. For example, a simplified "hidden object" mini-game is used to present a list of items found in an emergency Go-Bag. As the player progresses, pieces of badges are awarded. Each of the 17 badges represents an SDG.

To assist teachers with the integration of the game in their classes, we also provide a teacher guide. The guide contains a mapping of the various stories with the learning goals stipulated by the Philippines Department of Education. It contains recommended debriefing questions and personal projects that students can undertake to deepen learning.

4. Implementation of Nicholson's RECIPE

Gamification refers to the embedding of game elements such as points, badges, achievements, and leaderboards in non-game, day-to-day activities in order to motivate behavioral change. However, the use of extrinsic rewards, also known as rewards-based gamification, is often unable to produce lasting change. To encourage behaviors, therefore, Nicholson (2015) recommends the use of game design elements that increase intrinsic motivation or meaningful gamification. Meaningful gamification refers to the use of game elements that will enable users to connect a game experience with their own beliefs and allow the transformation of these beliefs, leading to lasting change.

Since the experience we are creating is a game in itself, it may therefore be said that a gamification framework is inappropriate because we are not in fact layering game elements, whether intrinsic or extrinsic, on an existing real-world system. Game-based learning, referring to the creation or use of a game to drive home a specific instructional message (Ingwersen, 2017), would be more appropriate. However, we found that Nicholson's (2015) six elements apply to games in general and not just gamification specifically. These elements are: Play, Exposition, Choice, Information, Engagement, and Reflection. When reordered, they spell the acronym RECIPE.

4.1 Play

Play refers to voluntary activities that occur within a defined space, e.g. a stadium or a field. Play has rules but also allows for exploration. Play has constraints but allows players to modify those constraints as they progress. Most critically, play is intrinsically rewarding and does not rely on external rewards to drive the activity.

For People and Planet is intended to be a supplement to classroom instruction. It is up to teachers and students whether to use the game or not. Once in the game, players can play stories in any order. They can save their games and continue where they left off. In order to complete each quest and obtain the badge pieces, players must play the mini-games at least once. However, the badge piece is awarded regardless of score. In the Waste Segregation mini-game, for example, the player has to put as much trash as possible in the correct bin within a 30-second time limit. Putting the trash in the correct bin increases the player's score but putting the trash in the wrong bin results in penalties. The player can play the game earnestly or can simply let the time run out. In either case, the player earns the plot item needed to complete the quest and the plot item stays in the player's inventory even if the player replays the story. This was a design choice made for players who would like to replay the story but not the mini-games. Players can choose to replay mini-games to achieve higher scores but these scores are not shared publicly, so this choice is wholly for their own personal satisfaction.

4.2 Exposition

The narrative layer of the game is the *Exposition*. Aside from serving as a backdrop for the game elements, the narrative, if reflective of reality, provides the player with a means of connecting to and understanding the real world. Brand and Knight (2005) identify four dimensions of narrative elements in games: the evoked narrative, the enacted narrative, the embedded narrative, and the emergent narrative. The evoked narrative refers to a pre-existing mythology or franchise in which the game is rooted. The enacted narrative is the delivery of the narrative through game elements and cut scenes. Embedded narratives are the backstories that the player discovers in the course of interacting with the game. Finally, the emergent narrative is the story that the player constructs by making meaningful choices while playing.

For People and Planet is an original work with no roots in prior mythology. However, the game scenarios are drawn in day-to-day life experiences, so we may argue that the embedded narrative is day-to-day life in the Philippines. The enacted narratives occur when each of the quests is completed. For example, once the player has helped the farmer Ate Chay, the player receives a basket of fruits and vegetables in a cutscene. The non-player character backstories provide the embedded narrative. In "A Walk in the Park", the player meets Tanya, a young girl with a disability. The player learns that Tanya was the victim of bullying and was wounded because of the incident. The player's decision to take Tanya to a clinic for first aid is an example of an emergent narrative.

4.3 Choice

This leads us to the next element: *Choice*. The player needs to have some control or autonomy over how he/she interacts with the system. In games for learning specifically, learners should ideally have a choice of activities from which they can learn the same concept. Regardless of their goals, learners have to be given guides so as to empower them to reach their goals.

As mentioned earlier, *For People and Planet* does not enforce a sequence of stories. The player is free to move around the game. Furthermore, the game provides multiple opportunities to learn about an SDG and the different paths to achieving it. As shown in Table 1, SDG #3 Good Health and Well-Being is covered in stories four of the five stories while SDG #14 Life Below Water is covered in "What's for Lunch" and "Work, Work".

Narratives and interactions with non-player characters provide the player with support about what to do next. When Ate Chay tells the player, "You see, it's those pests again. I usually get rid of them myself but my back hurts today. How am I supposed to get any work done now?" this cues a mini-game in which the player helps Ate Chay remove the pests from her crop. When the fisherman Manong Max says, "You're here for some fish, aren't you? If you help me with my fishing and ocean cleanup duties, then I'll give you some of today's catch." this leads to two mini-games: one to collect trash from the ocean and another to catch adult (as opposed to juvenile) fish.

4.4 Information

Information refers to game feedback that enables learning about the real-world context as well as the real-world consequences of their choices. Players can receive information through the visual displays that represent player or world statistics, through dialogs with non-player characters, through the narratives, or even through game mechanics.

For People and Planet integrates information about sustainability throughout the gameplay. The information can precede mini-games or be part of the game mechanics. Before Waste Segregation, for example, the player is told about the importance of segregating household waste. In Pest Control, the player uses a spray bottle with pesticide but the preceding narrative explains that pesticides are one of several ways in which farmers can control pests. In the Go Bag mini-game, the player is prompted to find items that should go into the bag. The items should indeed be part of a go bag such as flashlight, water, and some ready-to-eat food.

4.5 Engagement

Engagement in this context has two definitions: Social engagement which refers to cooperative, collaborative, or competitive interactions with other players that enable discovery and learning, and an engaging game experience in which game difficulty and player skill are well-matched.

For People and Planet is not a multiplayer game nor does it have a social media presence. However, it is intended to be played in a classroom setting or as part of the class. While the game is playable in itself, the intention was for it to be part of a teacher’s lesson. This opens the possibility of social engagement with and through the game. Within the game itself, social interaction is simulated through interactions with non-player characters. The player cooperates with friends and relatives in order to achieve the game’s goals.

4.6 Reflection

Finally, *Reflection* refers to activities that deepen engagement and learning by assisting participants to find other interests and past experiences that connect with the game. These opportunities usually occur in the form of debriefings that challenge learners to step back and think about the experience.

The game implements this by referring to actual incidents or social realities. In “Flood Fighters”, the player’s grandmother talks about the disaster wrought by Typhoon Yolanda in 2013 (See Figure 1). In “A Walk in the Park”, the narrative shares that people with disabilities are sometimes targets for harassment. Reflection is further encouraged through the recommended activities. Out-of-game reflection is encouraged with debriefing questions and personal projects. One such personal project for “A Walk in the Park” is to try to assess how disabled-friendly a place is. Students are asked, “Are there facilities such as easy-to-use ramps, elevators so that wheelchairs can access upper floors, or passable sidewalks and walkways? If there aren't enough, where do you think these can be added?”

5. Summary and Future Work

In this paper, we described the design and development of a game that teaches about the SDGs that is contextualized in the Philippines. We were motivated by the need for sustainability awareness that is rooted in Filipino scenarios. Nicholson’s (2015) RECIPE framework guided our design choices. In this paper, we give examples of how we implemented the principles of Play, Exposition, Choice, Information, Engagement, and Reflection in *For People and Planet*. The research team is currently completing the last few mini-games. In parallel, we are performing a user experience review and completing the teacher resource materials. Once the game and all ancillary materials are completed, we will we will conduct teacher training in the use of the game for classes and will perform a field test with our target learners.

Acknowledgements

We thank our support team composed of Ma. Rosario Madjos, Jose Isidro Beraquit, Rommel Ian De La Cruz, Fr. Johnny Go, SJ, Tomas Falgui, Hakeem Jimenes, Dan Mark Restoles, Toby Rubio, and Jaclyn

Ting Ting Vidal. This work was made possible through the grant entitled, “For People and Planet: Gamified Tutorials for Middle School on Sustainable Development” from the University Research Council of the Ateneo de Manila University.

References

Bevilacqua, M., Ciarapica, F. E., Mazzuto, G., & Paciarotti, C. (2015). “Cook &Teach”: learning by playing. *Journal of Cleaner Production*, *106*, 259-271.

Catan GmBH. (1996). *Settlers of Catan* [Board game]. Germany: Catan.

Centre for Systems Solutions. (2019). *The World’s Future*. Retrieved from https://worldsfuture.socialsimulations.org/

Filho, W. L., Shiel, C., Paço, A., Mifsud, M., Ávila, L. V., Brandli, L.L., Molthan-Hill, P., Pace, P., Azeiteiro, U.M., Vargas, V. R., Caeiro, S. (2019). Sustainable Development Goals and sustainability teaching at universities: Falling behind or getting ahead of the pack? *Journal of Cleaner Production. 232*, 285-294.

Hensley, N. (2020). Educating for sustainable development: Cultivating creativity through mindfulness. *Journal of Cleaner Production*, *243*, 118542.

Immacocollabo. (2018). 2030 SDG Game. Retrieved from https://2030sdgsgame.com/2030-sdgs-game/

Ingwersen, H. (2017). Gamification vs. game-based learning: What’s the difference? Retrieved from https://blog.capterra.com/gamification-vs-games-based-learning/

Katsaliaki, K., & Mustafee, N. (2012, December). A survey of serious games on sustainable development. In *Proceedings of the 2012 Winter Simulation Conference* (WSC) (pp. 1-13). IEEE.

Meadows, D.H., Meadows, D.L., & Behrens III., J. (1972). *The Limits to Growth*. Washington DC: Universe Books.

Nicholson, S. (2015). A recipe for meaningful gamification. *Gamification in education and business*, 1-20, Springer, Cham.

Nousheen, A., Zai, S. A. Y., Waseem, M., & Khan, S. A. (2020). Education for sustainable development (ESD): Effects of sustainability education on pre-service teachers’ attitude towards sustainable development (SD). *Journal of Cleaner Production*, *250*, 119537.

Philippine Business for the Environment (PBE). (2018). SDGs Our Biz. Report on PBE-UNDP project.

Prensky, M. (2006). Don't bother me, Mom, I'm learning! How computer and video games are preparing your kids for 21st century success and how you can help. *St. Paul: Paragon House*.

Purvis, B., Mao, Y., & Robinson, D. (2019). Three pillars of sustainability: in search of conceptual origins. *Sustainability science*, *14*(3), 681-695.

UN. (2015). The Millenium Development Goals Report. NY: United Nations. Retrieved from https://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20(July%201).pdf

UN. (2016). The Sustainable Development Goals Report. NY: United Nations. Retrieved from https://unstats.un.org/sdgs/report/2016/The%20Sustainable%20Development%20Goals%20Report%202016.pdf

UN. (2019). *Frieda Makes a Difference: The Sustainable Development Goals and How You Too Can Change the World*. NY: United Nations.

UN. (2019). *Go Goals! Welcome to the SDG Board Game for Children* [Board game]. Retrieved from https://go-goals.org/

We are Muesli (n.d.). Once upon a tile: A (not so) casual puzzle life-sim. Retrieved from https://www.wearemuesli.it/out/

Willamo, R., Helenius, L., Holmström, C., Haapanen, L., Sandström, V., Huotari, E., Kaarre, E., Värre, U., Nuotiomäki, A., Happonen, J., Kolehmainen, L. (2018). Learning how to understand complexity and deal with sustainability challenges – A framework for a comprehensive approach and its application in university education. *Ecological Modelling. 370*, 1-13.

Zu Digital. (2020). *World Rescue*. Retrieved from http://worldrescuegame.com/

Designing Games for Stealth Health & Healthy Lifestyle Education

Nilufar BAGHAEI[a*], Ilona HALIM[a], John CASEY[b], Samantha MARSH[c] & Ralph MADDISON[c,d]
[a]*Games and Extended Reality Lab, Massey University, Auckland, New Zealand*
[b]*School of Computing, Unitec Institute of Technology, Auckland, New Zealand*
[c]*National Institute for Health Innovation, The University of Auckland, New Zealand*
[d]*Institute for Physical Activity and Nutrition, Deakin University, Australia*
*n.baghaei@massey.ac.nz

Abstract: Mobile games can be highly accessible and effective tools in educating and promoting children's health education, given the significant increase in their popularity. In this paper, we describe our experience in designing mobile games for enhancing children's knowledge of healthy diet and lifestyle. We discuss design strategies for embedding lifestyle educational content and engagement with healthcare providers into three distinct mobile games. Our findings show that the children found the games engaging and enjoyed interacting with them. This work makes an important contribution to the field of games for health. It aims to adopt a 'stealth health' approach to engaging people in their own health care management by leveraging a technology that is currently used and accepted by the target population.

Keywords: Game design, mobile game, lifestyle, games for health, stealth health

1. Introduction

The prevalence of type 2 diabetes (T2D) in New Zealand has been increasing at an alarming rate, and it is a public health priority that needs to be addressed. T2D is a leading cause of death for New Zealanders, including Maori (indigenous population of NZ) (Health, n.d.). The disease is associated with other comorbidity including renal failure, lower limb amputation, avoidable vision loss and blindness and heart disease; it is also a major contributor to the inequalities in life expectancy (Ministry of Health, 2008).

Existing research evidence provides support for the use of videogames to promote health-related behaviours (Baghaei et al., 2016). Adopting a healthy lifestyle can reduce the chance of children developing T2D. A randomized controlled trial, which provided information and behaviour change support/strategies in the form of a videogame, helped increase fruit and vegetable consumption by 0.67 servings per day ($p<0.018$) in children aged 10–12 years (Baranowski, et al., 2011). A systematic review of 14 randomised controlled trials of interactive multimedia interventions to promote communication of dietetic messages with overweight pre-adolescent children demonstrated potential to improve children's health-related self-efficacy, which could in turn enable them to become more competent on complex topics such as dietary behaviour change discussions (Raaff, Glazebrook, & Wharrad, 2014). It also highlighted potential of multimedia interventions to support communication between young children and health professionals.

In recent years, designing, developing and playing games with smartphones and tablets have become increasingly popular both in research and industry. The popularity and ubiquitous use of smartphones and tablet computers offers considerable potential to deliver interventions that would support communication between young people and health professionals. For example, smartphone-based games could be played while waiting in outpatient clinics or between visits with healthcare professionals (Greysen et al., 2014).

According to Kirriemuir (2002), there are two key themes common to the development of games for education: (1) the desire to harness the motivational power of games to "make learning fun"; and (2) a belief that "learning through doing" in the form of games offers a powerful learning experience. It has been proposed that the real educational value of a computer game should be exemplified by its ability to create a playful learning experience for children through experimentation, progressive exploration, trial and error, imagination, role play, and simulation. Therefore, a game designed to satisfy these criteria might provide a useful platform for education. Based on this understanding, learning in a computer game should be purposely structured through a series of exploration tasks so that children can discover essential diabetes knowledge in a progressive and experimental manner. Driven by an initiative from the Adult & Pediatric Diabetes Psychology Service of New Zealand, research was undertaken to design and develop effective approaches for lifestyle education and enhancing children's engagement with healthcare provider. In this paper, we describe the process of designing several health educational prototype mobile games for increasing children's knowledge of healthy diet and lifestyle and to encourage them to engage with their healthcare provider on a regular basis.

2. Game Design for Lifestyle Education

The aim of our project was to determine how to use video games to enhance children's knowledge of healthy lifestyle and their engagement with healthcare providers. The research questions we investigated in this paper was whether participants enjoyed playing the games and had a positive perception.

2.1 Design Strategies

The key concept that is frequently utilized to explain the level of engagement in a computer game is that of "flow", first introduced by Csikszentmihalyi (1990). Many researchers consider flow as the state of intensive involvement. It is widely believed that flow is the key to the success of an educational game. According to Malone (1980), several conditions are likely to induce the flow state. Among them, three conditions are of particular importance for designing diabetes education games (Chen, Baghaei, & Sarrafzadeh, 2011):

- C1: The game should be designed so that it has levels of difficulty that can be adjusted to match the children's current diabetes knowledge.
- C2: The game should provide output to children that gives feedback on how well they performed on a particular activity and how to improve their performance. In this, the activity should be designed so that reasoning behind children's decision making in the game mimics the situation often presented when managing their diabetic condition in real life.
- C3: In game activities should include a variety of challenges that includes a message about different aspects of diabetes management, each with increasing level of complexity, so that children can learn a wider range of information about managing their condition.

To apply the educational features into the game, we used design strategies proposed in our previous work (Baghaei et al., 2011, 2016). The first design strategy was Structure Enhancement (SE), which means that the educational content should enhance the structure of the game and not weaken it. For instance, the addition of educational elements into the game should prompt design for a new series of difficulty levels so that children can gain knowledge progressively as they advance through the stages of the game. The second strategy was Feedback Enhancement (FE). With this strategy, the educational features must provide a knowledge-rich visual feedback that are triggered by specific in game situation. Feedback can be provided in various forms, such as message boxes and on-screen performance indicators. For example, a warning message can be shown to the user or have the performance indicator flash red when their in-game character chooses to eat unhealthy food. The final strategy is referred to as the Challenge Enhancement (CE). Our previous study suggested that challenge induced proactive knowledge discovery and encouraged engaged learning. Hence, CE recommends that educational features in a game should provide users with a range of challenges as they play the game. Using a stage-based game, simple educational features can be used for the early stages of the game, while complex features are reserved for more advanced stages of the game.

When designing an enjoyable educational game, it was also noted one of the most important aspect to consider was player's perception on the game. We used four heuristics used to measure player's perception on the game in the design and development of the games: Usability, Educability, Mobility, and Playability (Baghaei et al, 2016). These heuristics has worked well as a measurement of participant's experience of the game for the past diabetes educational game based on Mario Brothers, hence we continued to keep these heuristics in mind when designing the video games for this study.

2.2 Modifications of Ari & Friends

When choosing the game design, we initially chose to use the Mario Brothers game as a template for several reasons (Baghaei et al., 2016). Overall, we considered the Mario Brothers game well suited to provide a lasting and fun learning experience. Based on these strategies, additional features were added to the original open source Mario Brothers game to create 'Ari & Friends'. In this game, in addition to the original storyline where the main character has to explore various stages to find and rescue a princess, players are presented with an additional challenge where Ari is presumed to be suffering from T2D. As a result, players must solve health problems caused by the condition, such as the necessity to maintain therapeutic blood sugar levels. While exploring dungeons, players need to carefully consider the actions they take, as movements such as jumping, running and picking or eating food items will affect their blood sugar level, which are reflected in an indicator bar on the screen (Figure 1).

In the game, players are encouraged to eat healthy food to maintain energy and to keep active by exploring the terrain. It was expected that children would be able to make the connection and learn the skills to keep themselves healthy by choosing a healthy diet, being physically active, and maintain their blood sugar level. To encourage children to approach health care professionals when they needed assistance, a new character "doctor" was introduced to this version of the game. When players reach a checkpoint, they would encounter the doctor who asked health related questions. Players were rewarded with physical immunity against enemies and free adjustment to their blood sugar level if required. All modifications were consistent with the SE, FE and CE design strategies. An example of the use of these strategies:_As with the initial version of Ari & Friends, the main challenge was to maintain Ari's blood sugar level throughout every stage of the game. Players were required to keep a close eye on the indicator and maintain levels within predefined limits. If the blood sugar was low, players were required to use their remaining energy to find and eat a food item. If it was too high, they were required to undertake activities and exercise such as running and jumping. With these modifications, players were now required to explore the terrain to find food items and carefully choose which food to consume. This approach with SE strategy as the need to make correct food choices and being mindful of blood sugar level provides fine-grained challenges through every stage of the game.

2.3 Designing 3D Diabetes Educational Game

Using the same design principles and strategies, we also developed a 3D education game targeting a similar demographic. Similar to Ari & Friends, the 3D Zombie game intended to teach children about healthy lifestyles and management of T2DM, and to encourage them to engage with their healthcare provider on a regular basis. While the educational messages were similar, the game design was markedly different. 3D Zombie game had mechanics like Microsoft Minecraft, a game that is vastly popular among our target demographic. This provided an advantage of a shorter learning curve required to play the educational game. For the development process, we decided to use the Unity Game Engine to facilitate the rapid development of a fully 3D immersive world. The game was a mission-based

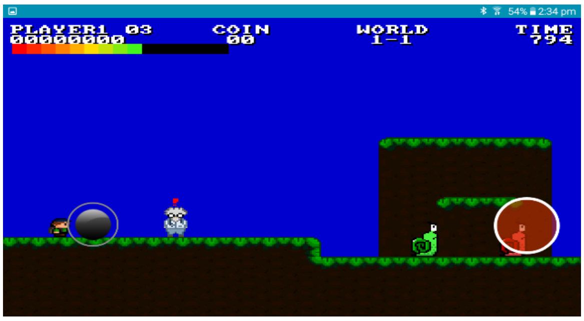
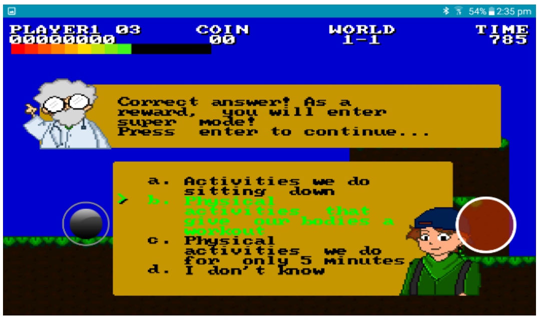


Figure 1. Ari & Friends Game Screen with Indicator Bar.



Correct answer! As a reward, you will enter super mode! Press enter to continue...

a. Activities we do sitting down
> b. Physical activities that give our bodies a workout
c. Physical activities we do for only 5 minutes
d. I don't know

Figure 2. Option Text Turns Green indicating Correct Answer.

role-playing game where players roamed a post-apocalyptic world overridden by zombies. Users played as a main character with T2D. Players were given tasks to complete in order to advance to the next level; each task related to a particular aspect of diabetes management. Players were required to find hidden healthcare providers (e.g., doctors or nurses) that could offer instructions or tips on how to solve a task and advance to the next level. If players ignored or ran away from a healthcare provider, they lost points, experienced changes in blood sugar level, or their movement speed slowed down (Figure 3).

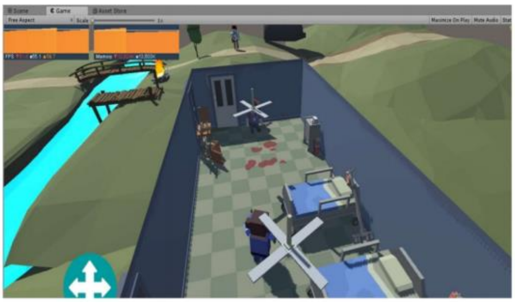
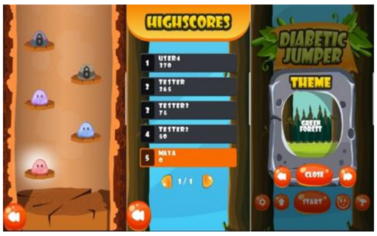


Figure 3. Screenshot of the Zombie 3D Game; the player wakes up in the hospital and must save others by following instructions of health care providers.

The player's main task was to save a person from turning into a zombie by completing healthcare provider's instructions. At the end of each level, a quiz session was triggered to assess player's knowledge and to collect data. The session comprised questions related to the tasks on each level, which made each task a learning goal and highlighted the need to complete every available task to advance to the next level. Similar to Ari & Friends game, one of the main challenges in the game was maintaining therapeutic blood sugar level. Because of the main character's condition, blood sugar level had a major impact on the character's performance. Different types of food items were scattered throughout the terrain; each resulted in different effects when consumed (i.e. speed increase) and affected blood sugar level differently. We initially included guns as a potential weapon for players to use against zombies. However, it was decided to forego the guns after talking to our potential participants and further consideration about our target age group. In the place of guns, we added a shield feature so that the players could defend themselves against zombies and sugary food items. We conducted a pilot study (n=10) during which we decided this game, with the current implementation, was not suitable to be played on a tablet or a mobile device and was aimed more for PC players. The game also had some performance issues. We decided not to go ahead with it, despite the visual appeal.

2.4 Designing Diabetic Jumper

Diabetic Jumper was a 2D game with the mechanism based on another popular video game, Doodle Jump (https://en.wikipedia.org/wiki/Doodle_Jump). The goal of this educational game was to improve children's health literacy and knowledge about diabetes management, and to give positive reinforcement for children to engage with their healthcare provider on a more regular basis. The players must jump on platforms to reach the finish line of each level to advance to the next stage. Players navigate the avatar's movement trajectory by tilting their devices left or right. There were several playable characters within the game. Characters were drawn as imaginary non-human avatars with bodies that has a condition similar to somebody with type 2 diabetes, so they must strive to maintain therapeutic blood sugar level. The game was comprised of six stages with increasing complexity. After some feedback from the pilot study, different game elements were introduced in this game in addition to blood sugar level, such as body weight and sleep level that needs to be maintained in the more advanced stages. This provided fine-grained challenges to the existing game structure, following the SE design strategy. In their journey to the top, they will encounter different items of various types, each type has different characteristics affecting either blood sugar, body weight, or sleepiness. In the later version of the game, items can also have other interesting effects, such as gradually increasing weight, gradually decreasing blood sugar level, and instant death.



1	USER4 370
2	TESTER 265
3	TESTER3 75
4	TESTER2 50
5	MAYA 0

Figure 4. Some of the Additional Features of the Game.

The SE, FE and CE design strategies were used in the design process of the game. Pickable item types started with modest variation in earlier stages with types of healthy or unhealthy food, water, and exercise items. As the player unlocks more advanced stages, item types will have more variation, gradually adding more situational problems for the players to solve, such as drowsiness that begins from the medium difficulty stages and passing food carts with junk food that will kill the players instantly in more advanced stages. This way, children can first learn how to maintain the condition by using simple food choices and exercise, then learn how to maintain increasingly complex conditions and learn how to make more complex decision making as work their way up to later stages. This aligns with the CE design strategy where challenges are structured so that children can discover and learn educational content in a progressive manner. To give positive reinforcement on engaging with healthcare professional, we introduced checkpoints in each stage where the character can rest and answer quizzes given by a nurse character. The quiz screen displays a question with multiple answer choices.

The game followed the FE design principle by giving clear visual feedback after the player has given an answer. The choice will be highlighted in green for correct answer and red for incorrect answer. Further information related to each question will be displayed regardless of correct or incorrect answer. However, players are rewarded with point bonuses and free stats adjustment (blood sugar level, weight level, and sleep level when applicable) if they answer correctly. Visual feedback are also given in the form of explicit colour change in the indicator bars of blood sugar level, weight, and sleep level when they enter unsafe or critical level, and green in therapeutic level. When an indicator enters an unsafe level, a warning message will pop up, warning the user and giving instructions on how to bring it back to therapeutic levels. In addition to the gameplay, the application has features such as level selection, game settings, help, and personal high scores (Figure 4) to improve Usability and Playability.

3. Pilot Study

A series of co-design user-panel workshops with young people aged 9-15 years were conducted. During these workshops, young people played the prototype games and then provided user feedback directly to the designers and developers. These interactive sessions allowed users to highlight factors that facilitated or inhibited game use. They also provided detailed input into possible changes to improve the games. Based on feedback, the game developers revised the game and the process was repeated until we had a working version of the game to pilot test. In total, three user panels were conducted over a period of 12 months. Further, having revised the games we beta-tested with 10 participants aged 9-15 years over one week. After one week, all participants said that they played the games regularly over the course of 7 days. After allowing some time for the participants to experience the gameplay of all three games, we questioned the participants on their preference between the three games and what they thought about the games. The majority of the participants voted for the Diabetic Jumper as their preferred game, with only one participant reporting they preferred the 3D game. Participants stated the game controls were difficult to use on a tablet and would prefer to play it on a PC.

The visualisation of the bubble for choosing items in Diabetic Jumper was well received by the participants. During the session, participants also made some suggestions on how to improve the game both in terms of engagement and educational value, e.g. adding themes such as Halloween, Christmas, School party with different food items and backgrounds, introducing factors that immediately kill the player such as a fast food restaurant, including exercise bubble and pillow for sleep, making the levels more challenging as the players progress through the game. As a result of the Beta testing, we decided to use two of the games only (Diabetic Jumper and Ari & Friends), which were tested in a small pilot study of potential end-users.

4. Conclusion and Future Work

This project proposed novel design ideas for delivering healthy lifestyle education and interactions with health care professionals. Using co-design principles, we designed and developed three prototype serious mobile games for health targeting children. We further developed expertise and skills in the use of co-design methodology applied to the design and development of serious games for health. These methods can also be applied to other health conditions. The games are currently suitable for Android devices and aim to promote knowledge about healthy lifestyles and to engage young people to discuss their health condition with their healthcare provider or other similar person. It is possible these games could be used in a clinic or general practice setting to encourage young people to discuss their health condition. Initial findings showed that the children found the games engaging and enjoyed interacting with them. Future research is required to test the potential of these games in a primary care setting and examine whether children's knowledge of healthy lifestyle will enhance as a result of engaging with these games over a period of time.

This work makes an important contribution to the field of serious games for stealth health and healthy lifestyle education. Our approach aims to adopt a 'stealth health' approach to engaging people in their own health care management by leveraging a technology that is currently used and accepted by the target population.

Acknowledgements

We would like to thank the Health Research Council of New Zealand for funding this project. We would also like to thank Rinki Murphy, Rosie Dobson and Amanda Calder for their contributions.

References

Baghaei, N., Nandigam, D., Casey, J., Direito, A., & Maddison, R. (2016). Diabetic Mario: Designing and Evaluating Mobile Games for Diabetes Education. *GAMES FOR HEALTH JOURNAL: Research, Development, and Clinical Applications , 5*(4).

Baranowski, T., Baranowski, J., Thompson, D., Buday, R., Jago, R., Griffith, M. J., . . . Watson, K. B. (2011). Videogame play, child diet, and physical activity behavior change:A randomized clinical trial. *American journal of preventive medicine, 40*(1), 33-38. doi:10.1016/j.amepre.2010.09.029

Chai, C. S., Koh, E., Lim, C. P., & Tsai, C.-C. (2014). Deepening ICT integration through multilevel design of technological pedagogical content knowledge. *Journal of Computers in Education*(1), 1-17. doi:10.1007/s40692-014-0002-1

Chen, G., Baghaei, N., & Sarrafzadeh, A. (2011). Designing Games to Educate Diabetic Children. *Australian Computer-Human Interaction Conference.* Canberra.

Cziksentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience.* New York: Harper & Row.

Greysen, S. R., Khanna, R. R., Jacolbia, R., Lee, H. M., & Auerbach, A. D. (2014). Tablet computers for hospitalized patients: a pilot study to improve inpatient engagement. *Journal of hospital medicine, 9*(6), 396–399. doi:10.1002/jhm.2169

Kirriemuir, J. A. (2002). *Survey of the use of computer and video games in classrooms*. (Internal Report for British Educational Communications and Technology Agency) Retrieved from http://www.digra.org/wp-content/uploads/digital-library/ 05150.28025.pd

Malone, T. W. (1980). What Makes Things Fun to Learn? A Study of Intrinsically Motivating Computer Games. *Palo Alto: Xerox*.

Ministry of Health. (2008). *Diabetes and Cardiovascular Disease Quality Improvement Plan.* https://www.health.govt.nz/system/files/documents/publications/diabetes-cardio-quality-improvement-plan-feb08-v2.pdf, accessed May 2021

Raaff, C., Glazebrook, C., & Wharrad, H. (2014). A systematic review of interactive multimedia interventions to promote children's communication with health professionals: implications for communicating with overweight children. *BMC medical informatics and decision making, 14*, 8. doi:10.1186/1472-6947-14-8

Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Children Preference Analysis of a Logic Game- "Lily's Closet"

Wei Tung NIEN, Yi Chen WANG & Joni Tzuchen TANG*
Graduate Institute of Applied Science and Technology, National Taiwan University of Science and Technology, Taiwan
*jttang0@mail.ntust.edu.tw

Abstract: Children autonomously play games. It is what Game-Based Learning is aiming for, that to create a learning process for children to acquire knowledge autonomously. This study is to discover the relations among children's age, preference, and learning performances in a digital game, Lily's closet. The researchers applied big data analysis and collect gaming behavior data from 6,924 children whose age range is 3-8 years old. The result shows that: a) Children's age is related to learning performance. The higher ranking shows higher learning effects. The relevance between ranking and age shows that the older children grow up, the better "logic concept" they can comprehend. b) However, the older they are, the more toys and environment they are approaching. That is why their willingness to repeat the same game is lower after growing up. As a result, we suggest that to assist learning, we should give children a different level of games according to their ages. The research hopes to resonate with readers and jointly explore the relationship between adaptive development and digital games.

Keywords: Digital games, game-based learning, big data analysis, autonomously play, adaptive development, learning performance

1. Introduction

"Games Generation," which is also called "G Generation," is a group of people who grow up by the sound and light stimulation of the game (Prensky, 2001; Gibson et al., 2007). Many studies showed children keep playing games is that games are interesting, exciting, and challenging in problem-solving (Gerkushenko & Gerkushenko, 2014; Dalton & Devitt, 2016). In-game scenarios, children "enjoy" the stimulus if they can feel they are capable of coping with challenges (Gerkushenko & Gerkushenko, 2014; Dalton & Devitt, 2016).

When we know that games are very important for young children, the development of logic concepts has always been an important milestone in the development of young children. We need to understand: 1. Children of different ages have different learning needs. Is there any trend in logic concepts learning in games? 2. Does one logic game have the same appeal to children of different ages? In response to the above problems, our research purpose is to discover the trend of children's digital play-based learning. Our study questions are as follows:

1. What is the relation between age and gaming frequency?
2. What is the relation between age and the learning effect?
3. What is the difference of children's playing behaviors and the learning effect between kids who are highly involved in the game and kids who are not?

2. Theoretical Structure

2.1 The significance of Digital Games to Children's Autonomous Learning

Digital Game-based Learning (DGBL) is a kind of gameplay that defines the learning outcomes and it involves the game rule design. The idea of GBL believes that if we can motivate children and allow them to develop a learning awareness, children can automatically learn and obtain knowledge and information (Van Eck, 2006).

DGBL can improve children's attention and learning motivation, and children's learning effect is better in DGBL (Käser et al., 2013; Gerkushenko & Gerkushenko, 2014; Godwin et al., 2015; Mouws & Bleumers, 2015; Dalton & Devitt, 2016; Aunio & Mononen, 2018). However, the digital tool is a double-edged sword. If a digital game is not designed for learning purposes, and mediators like parents and teachers do not teach children how to use the digital product correctly, games may hurt children's cognitive development or cause wrong cognition (Lieberman et al., 2009). DGBL is an irreversible trend, therefore, to design a well-designed game that assists children's development, hobbies, and ability is an important task for future human beings.

This research creates a situation where children can play games autonomously. Research allows children to freely choose games in their own free time. Therefore, this research uses big data to observe the frequency of children's willingness to play games independently, and explores the relationship between different ages, genders, and learning effectiveness through research questions.

2.2 The Connection between Children's Logic Concepts Development and Game Design

The game in our research is about children's early logic concspts learning. Our study is based on Piaget's Cognitive-Developmental Theory. *Cognitive-Developmental Theory* explains how children develop their realization to the outer world, and the basic element of the development is called schema (Piaget, 1952; Hunting, 2010). Piaget conveyed that the establishment of cognition is an active process (Crain, 2014; Herczeg et al., 2019) based on prior knowledge (Bhattacharjee, 2015). Children meet cognition conflicts while interacting with the world and then build the new cognition through adaptation of the conflict by assimilation and accommodation (Piaget, 1962).

Our research participants are children in the Preoperational Stage between 3-8 years old. Children's cognitive thinking during the Preoperational Stage is fixed and concrete; they can read simple symbols but cannot think logically. While their cognition is developing to the next stage, the Concrete Operational Stage, they can do reasoning thinking according to the concrete stuff. Games are the way children learn new things (Piaget & Inhelder, 1969). There are various ways of gameplay, such as symbolic play or pretend play (Edna & Ronny, 2015), differentiate reality and fantasy (Woolley & Ghossainy, 2013), role-play, personification in the story, and perspective-taking (Yadi, 2020). Piaget considered that children would accommodate themselves in-game through transformation function, assimilate the new experience, and gain the knowledge. Children of different ages change their play contents and form along with the Different Stages of Cognitive Development, so as they constantly self-construct and develop in play (Piaget & Inhelder, 1969). Play material is a source of acquiring logic concepts for young children as well (Hunting, 2010; Piaget, 1962). Piaget (1962) stated that children construct logic concepts through hands-on play to explore and develop the realization of the object and logic relationship. Individuals involved in activities to acquire knowledge, then accumulate and build cognitive concepts step by step (Piaget, 1962). As a result, to children, the operation and repetition of games are important behaviors while learning logic concepts.

2.3 Meaning of "The Flow State" in Early Childhood

"Flow State" means a fully mental-involved activity. It is also called "being in the zone", which means players are in a mental state of fully immersed in a feeling of energized focus and enjoyment in the process of the activity (Csikszentmihalyi, 1998 & 2000; Cherry, 2018; Nakamura and Csikszentmihalyi, 2002). In other words, "Flow State" is characterized by complete absorption in what one does and a resulting loss in one's sense of space and time. Nakamura and Csikszentmihalyi (2002) identify six factors as experiencing the flow state:

- Intense and focused concentration on the present moment.
- Merging of action and awareness.
- A loss of reflective self-consciousness.
- A sense of personal control or agency over the situation or activity.
- A distortion of temporal experience, one's subjective experience of time is altered.
- Experience of the activity as intrinsically rewarding.

As a well-designed game can provide players' flow state (Csilszentmihalyi, 1998 & 2000; Cherry, 2018; Nakamura and Csikszentmihalyi, 2002), in this study, we employed DGBL to track the behavior of game players to see in which age group this game can stimulate their Flow State. Besides, we applied the Flow State Theory in the game design of the study that suggested challenges that match children's development. We observed children’s reactions to the game and measured their attention level to find their preference for it (Csikszentmihalyi, 1998 & 2000; Cherry, 2018; Nakamura & Csikszentmihalyi, 2002).

If a game stimulates children’s Flow State, they will spend more time and attention on such a game and be willing to choose it again and again (Csikszentmihalyi, 1998 & 2000; Cherry, 2018; Nakamura & Csikszentmihalyi, 2002). Therefore, to observe their Flow State, we observed the data of children’s behavior and analyzed their tendency according to their frequency in the game, so that we could see if the Flow State is activated.

3. Method

3.1 Equipment and Game

The logic game in this study is “Lily's closet.” (Figure 1)

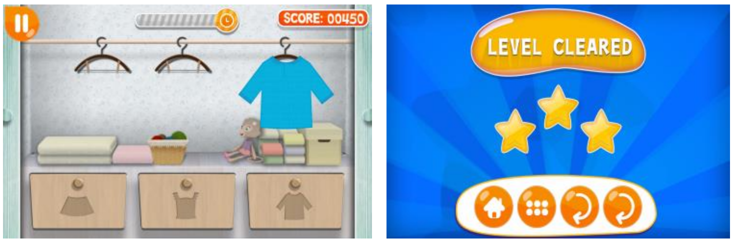


Figure 1. Screenshot of “Lily’s Closet”

“Lily’s closet” is a game about “identification and categorization” belonging to “logic” ability. In the game, users should touch and drag the right clothes shape to the matching closet. Once matched, users will score and the pad will give happy sound feedback. If the match is wrong, a hint will pop up with a warning sound. The game will give star levels after each session according to their gaming performance. Whenever players enter a game, the system starts to record automatically. The recorded content includes data such as entry time, game time, game level, stars scored (the lowest: 1; the highest: 3), user’s age, etc.

The definition of the key data we mainly analyzed in this study as follows.

- “Game Time”: means the length of gaming from entry to leaving the game. If the user taps into the game multiple times, the game time for each entry is recorded separately.
- “Frequency”: means how many times the user logs into this game. No matter the length of the game time, every log-in is counted as “Frequency: 1.”
- “Learning Effect” means the Stars that the user score. The lowest is 1 Star, and the highest is 3 Stars. This study hopes this game can improve children's logical ability of categorization/collation, and the learning effect will present as the star they scored.

3.2 Participants

Our research focused on children of 3-8 years old. We collected 6,924 children and 53,149 data. Then we divided five age groups (see Table 1). To compare the play frequency and their learning effect, we chose the top 100 children in each age group by their frequencies of the gaming data. Therefore, there are a total of 500 children who are the highest frequency among all in this game.

Table 1. *Number of Participants and Amount of Data*

Age Group(Year Old)	3-4	4-5	5-6	6-7	7-8	Total

Number of All Users	1929	1865	1557	1034	539	6924
Amount of All Data	17607	13984	10778	7300	3480	53194
Number of Highly Gaming Users	100	100	100	100	100	500
Amount of Data of Highly Gaming Users	5898	4622	4000	3387	2289	20195

4. Analysis Results

4.1 Frequency of DGBL in Different Age Groups

Average Frequency in Each Age Group shows in Table 2.

Table 2. *Average Frequency in Each Age Group*

Age Group(Year Old)	3-4	4-5	5-6	6-7	7-8	Total
All Users' Average Frequence	9.1	7.5	6.9	7.1	6.5	7.42
Highly Gaming Users' Average Frequence	59.0	46.2	40.0	33.9	22.9	40.4
Highly Gaming Users' Average Frequency ÷ All Users' Average Frequence	6.4	6.1	5.8	4.7	3.5	5.4

Unit: times

All children's average frequency is 6-9, while Highly Gaming Users' is 22-29. The difference between both is 3.5-5.4 times (Table 2). It represents that Lily's Closet can create high motivation in playing for children whose age range is 3-8 years old. They would want to play it over and over.

The Correlation Analysis between Age Group and Frequency shows in Table 3.

Table 3. *Correlation Analysis between Age Group and Frequency*

		Age Group	Game Time
Age Group	Pearson Correlation	1	-.314**
	Sig. (2-tailed)		.000
	Total Amount of Data of Highly Gaming User	20,195	20,195

**. Correlation is significant at the 0.01 level (2-tailed)

According to analysis (Table 3), the relation between age group and frequency has a significant difference as a negative correlation. Older children played fewer times than younger age.

4.2 Learning Effect of DGBL in Different Age Groups

This study recorded stars scored in each age group between 3-8 years old. We compared stars of Highly Gaming Users among each age group. Table 4 shows the average stars of all children and Highly Gaming Users. We analyzed the relationship among age groups and stars as Table 4 and 5 show.

Table 4. *Average Stars in Each Age Group*

Age Group(Year Old)	3-4	4-5	5-6	6-7	7-8
All Users' Average Stars	1.60	1.88	2.15	2.36	2.43
Highly Gaming Users' Average Stars	1.64	2.00	2.23	2.38	2.47

Table 5. *Correlation Analysis between Age Group and Stars*

		Age Group	Game Time
Age Group	Pearson Correlation	1	.375**
	Sig. (2-tailed)		.000
	Total Amount of Data of Highly Gaming User	20,195	20,195

**. Correlation is significant at the 0.01 level (2-tailed)

All children's average stars are 1.60~2.43, while Highly Gaming Users' are 1.64~2.47. Highly Gaming Users' average stars are all higher than all children's average. It represents that repeated gaming can help children to identify colors and shapes, and improve their categorization ability which is under a logic field. According to analysis, the relation between age group and stars has a significant difference as a positive correlation. Older children scored a higher rank of stars than a younger age.

5. Discussion

We would like to explore children's Digital Game-based Learning (DGBL) and their preference in this study. We discuss and answer the study questions as follows:

1. What is the relation between age and gaming frequency?

According to the analysis, the negative correlation between frequency and age group represents that Lily's Closet is more attractive to younger children, who have higher Flow State than older children. The factors that affect children's repetitive playing could be numerous, such as operating fluency of the game commands, comprehension of the rules, and challenging degree to them to evoke their motivation. If a game is too easy, children will end up not playing it because it cannot evoke a sense of challenge, though it may attract them first with sound and light effects. Children's logic ability is developed by the impact of social culture context, and they absorb the relative information through the living environment and contextual interaction (Piaget, 1962). Therefore, we convey that because older children own more life experience than younger children, it is less challenging for them to collate colors and shapes. As they can score three stars easily, they will not feel interested in it anymore.

2. What is the relation between age and the learning effect?

Children's age and their learning effect are significant relations. The analysis shows that the age group and average stars are positive correlations. It means that a game like Lily's Closet that classifies colors, shapes, and patterns, is attractive to children who are 3-4 years old. As they grow older, children's identification and classification abilities develop better, so that their stars and learning effects are higher.

3. What is the difference of children's playing behaviors and the learning effect between kids who are highly involved in the game and kids who are not?

All Highly Gaming Users scored higher stars than all children's average stars in any age group. We interpreted it as Highly Gaming Users' learning effect is better than average. According to the analysis, we suggested that playing Lily's Closet repeatedly can be a great practice for young children to improve their identification and categorization abilities. As Piaget (1962) mentioned, young children's cognitive process of constructing knowledge is through constant conflicts of concepts, adjustments, and then acceptance, so that they can form the cognition to the outer world. That is what so-called assimilation and accommodation process. The game rule of scoring in Lily's Closet is to touch and drag the clothes to the appointed closet. Once the wrong match happens, there will be a hint and red flashlight with a warning sound. Therefore, we consider that if children can repeatedly play DGBL games which improve cognition development, scored or not, the stimulus of the game can still help children to build their cognition right. The analysis data echoes our beliefs in this study.

Reflecting on the research results back to the theoretical framework, we found that:

1. The theory shows that digital games produce children's autonomous learning (Van Eck, 2006). Children are willing to play the game autonomously, and the children who are most willing to play autonomously for this game are 3-4 years old. The learning achievement of children with high autonomous play shows higher scores than the children with low autonomous play.

2. Logic games are constructed in cognitive schemes (Piaget, 1952; Hunting, 2010). This research shows that high-frequency gamers use this game to build cognitive schemes and improve their logical effectiveness.

3. Children enter the flow experience will continue to play the same game over and over again (Csilszentmihalyi, 1998 & 2000; Cherry, 2018; Nakamura and Csikszentmihalyi, 2002), which also reflect to our study that high-frequency players have the flow experience.

References

Aunio, P., & Mononen, R. (2018). The effects of educational computer game on low-performing children's early numeracy skills–an intervention study in a preschool setting. *European Journal of Special Needs Education*, 33(5), 677–691.

Bhattacharjee, J. (2015). Constructivist Approach to Learning – An Effective Approach of Teaching Learning. *International Research Journal of Interdisciplinary & Multidisciplinary Studies*, 1(6), 65-74.

Cherry, K. (2018). 'Flow' Can Help You Achieve Goals Understanding the Psychology of Flow.*Verywell Mind.*

Crain, W. (2014). *Theories of Development Concepts and Applications* (pp.123). UK: Pearson Education Limited.

Csikszentmihalyi, M. (1998). *Finding Flow: The Psychology of Engagement With Everyday Life* (pp.47). New York, NY: Basic Books.

Csikszentmihalyi, M. (2000). *Beyond Boredom and Anxiety: Experiencing Flow in Work and Play*. San Francisco: Jossey-Bass.

Dalton, G., & Devitt, A (2016). Gaeilge Gaming: Assessing how games can help children to learn Irish. *International Journal of Game-based Learning*, 6(4), 22–38.

Edna, O. & Ronny, G. (2015). Symbolic play and language development. *Infant Behavior and Development*, 38(1), 147-161.

Gerkushenko, S., & Gerkushenko, G. (2014). The Play Theory and Computer Games Using in Early Childhood Education. *International Journal of Game-based Learning*, 4(3), 47–60.

Gibson, D., Halverson, W., & Riedel, E. (2007). Games and simulations in online learning: Research and Development Frameworks (pp.175-188). USA: IGI Global.

Godwin, K., Lomas, D., Koedinger, K., & Fisher, A. (2015). Monster Mischief: Designing a Video Game to Assess Selective Sustained Attention. *International Journal of Gaming and Computer-Mediated Simulations*, 7(4), 18–39.

Herczeg, M., Winkler, T., & Ohlei, A. (2019). Ambient learning spaces for school education. *iCERi 2019 Proc*, 1, 5116-5125.

Hunting, R. (2010). Little people, big play, and big mathematical ideas. In Shaping the future of mathematics education: Proceedings of the 33rd annual conference of the Mathematics Education Research Group of Australasia (pp. 727-730).

Käser, T., Baschera, G. M., Kohn, J., Kucian, K., Richtmann, V., Grond, U., & von Aster, M. (2013). Design and evaluation of the computer-based training program calcularis for enhancing numerical cognition. *Frontiers in Psychology*, 4, 89.

Lieberman D. E., Bramble D. M., Raichlen D. A., & Shea J. J. (2009) Brains, Brawn, and the Evolution of Human Endurance Running Capabilities. In: Grine, F. E., Fleagle, J. G., Leakey, R.E. (Eds), *The First Humans – Origin and Early Evolution of the Genus Homo. Vertebrate Paleobiology and Paleoanthropology*. Dordrecht: Springer.

Mouws, K. & Bleumers, M. (2015). Co-Creating Games with children: a case study. *International Journal of Gaming and Computer-Mediated Simulations*, 7(3), 22–43.

Nakamura, J., & Csikszentmihalyi, M. (2002). The Concept of Flow. In C. Snyder, & S. Lopez (Eds.), *Handbook of Positive Psychology* (pp. 89-05). New York, NY: Oxford University Press.

Piaget, J. (1952). Autobiography. In E. G. Boring, H. S. Langfeld, H. Werner, & B. M. Yerkes(Eds.), *A history of psychology in autobiography* (pp.237-256). Worcester: Clark University Press.

Piaget, J. (1962). *Play, dreams and imitation in childhood*, Norton.

Piaget, J., & Inhelder, B. (1969). *The Psychology of the Child*, New York, NY: Basic Books.

Prensky, M. (2001). *Digital game-based learning*. New York, NY: McGraw-Hill.

Van Eck, R. (2006). Digital Game-based Learning: It's not just the digital natives who are restless. *Educase Review*, 41(2), 1-16.

Woolley, J. D., & Ghossainy, M. E. (2013). Revisiting the fantasy–reality distinction: Children as naïve skeptics. *Child Development*, 84(5), 1496-1510.

Yadi, Z. (2020, July). A Study of Preschool Children's Second Language Acquisition From the Perspective of Piaget's Theory of Cognitive Development Stages—A Comparison between Raz and Oxford Reading Tree. In *2020 5th International Conference on Humanities Science and Society Development* (pp. 657-665). Atlantis Press.

Xiphias: Using a Multidimensional Approach towards Creating Meaningful Gamification-Based Badge Mechanics

Jonathan DL. CASANO[a*], Jenilyn L. AGAPITO[a] & Nicole Ann F. TOLOSA[ab]
[a]*Ateneo de Manila University, Philippines*
[b]*Samsung R&D Institute, Philippines*
*jcasano@ateneo.edu

Abstract: This paper shows the design and initial testing of three new Xiphias Badges -- Presence, Mastery, and Antifragility – based on the merging of the salient features from James Clear’s Behavior Change model (2016), Johann Hari’s Lost Connections model (2018), and Jordan Peterson’s recent interpretation of the Big Five model of Personality Traits (2007). This multidimensional approach is an attempt to cater to the multidimensionality of a user and aims to be a more universal gamification approach that taps into internal motivations. The badge mechanics were tested on 69 undergraduate students using a Low-Fidelity Gamified Tracker. The results of a survey that sought their insights on the utility of the badges showed their potential to be motivating factors in the classroom.

Keywords: Gamification, meaningful gamification, xiphias, online distance learning

1. Introduction

One of the early definitions of gamification in the context of education is the addition of game-like elements and mechanics to a learning process (Deterding et. al, 2011). This early definition had been carried out by teachers through incorporating the use of points (Neve et al, 2014), badges (Ibáñez et al., 2014; Neve et al, 2014) and leaderboards (PBL) in the conduct of their face-to-face classes. These implementations reported positive outcomes such as increased engagement (Hamari et al., 2014), small but significant increase in short-term test scores (Bakkes et al., 2012) and generally positive user experience (Garner et al., 2005). Over time, criticisms for PBL gamification grew stronger as researchers noted some adverse effects of doing PBL. For instance, there are reports of attrition and disengagement among the students who see themselves at the bottom 25% of the leaderboards (Christy & Fox, 2014). Hence, initiatives towards Meaningful Gamification were started, where, in contrast to the PBL framework, the game mechanics applied to the learning process seeks to activate internal motivation (Sailer & Homner, 2020). In recent years, Yu-kai Chou, the pioneer of the Octalysis Gamification Framework (Chou, 2019) observed that these meaningful gamification attempts fail to capture the multi-dimensionality and deeper-rooted motivations of players/students receiving the gamification intervention.

This work responds to Yu-kai Chou’s invitation by designing three new Xiphias badges namely, the Presence badge, the Antifragility badge and the Mastery badge whose mechanics were built around the interfusion of the salient features of James Clear’s Behavior Change model (2016), Johann Hari’s Lost Connections model (2018), and Jordan Peterson’s recent interpretation of the Big Five model (2007). Results from an initial testing of the badge mechanics on undergraduate classes are then presented. In light of the above, the paper tries to answer the following research questions: RQ1: How will Meaningful-Gamification badge mechanics look if designed using a multi-dimensional approach? and RQ2: To what extent did the new Xiphias badges affect the student learning experience?

2. Theoretical Framework

2.1 The Behavior Change Model

In a recent meta-review, Clear (2018) presented the Behavior Change Model (BCM) as a framework to understand and facilitate enduring behavior change among learners.

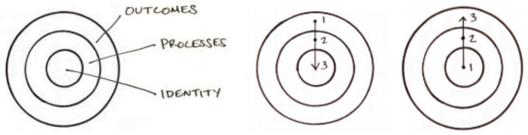


Figure 1. The layers of the Behavior Change Model (Clear, 2018).

In Clear’s model, an Outcomes-Based Intervention is one which focuses on using the Outcomes to motivate: “I need to pass the quiz (outcomes/motivation), so I will review for the quiz (process), and if I pass the quiz, I can say that I am a good student (identity)” – The motivation is from an external source. An identity-based intervention on the other hand is that which focuses on honing the Identity layer and using it to motivate: “I believe that I am a good student (identity), to keep this identity (motivation), I will review for the quiz (process) so I can pass the quiz (outcome)” – The motivation is internalized. The design of the new set of Xiphias Badges follows an Identity-based approach.

2.2 The Lost Connections Model

The Lost Connections Model of Hari (2019) posits that there are nine (9) common causes for Non-clinical Depression (NCD) among individuals (also cited in Song & Bonk, 2016). Among the nine common causes, four (4) are fixable through Social Prescription, or the method of merely exposing the individual to structured activities that allow the interaction with other individuals. These four Social-Prescription-fixable causes are (1) Disconnection from Meaningful Work, (2) Disconnection from Others, (3) Disconnection from Meaningful Values and (4) Disconnection from Status or Respect. The new Xiphias badges seek to incorporate the four Social-Prescription-fixable disconnects in its design.

2.3 The Big Five Personality Traits Model

The authors tried finding a model of personality whose categorizations are decently universal (Satow, 2021) and whose correlations to motivation are also well established and documented (John & John, 2020). The Big Five Personality Traits are namely Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A) and Neuroticism (N). A capacity for extended motivation among others may be classified and extrapolated from these five traits (Watson, 2019). We tried designing the new badges such that the mechanics would appeal to at least one extreme in the OCEAN spectrum.

2.4 The New Xiphias Badge Mechanics

Table 1. *The New Xiphias Badge Mechanics Vis-a-vis the Multidimensional Framework.*

Badge	Badge Mechanics	Behavior Change Layer (Clear, 2018)	Lost Connection Addressed (Hari, 2018)	Big Five Target Personality Trait (Peterson,2007)
Presence	See 3.4.1 paragraphs 1 and 2	Identity	Reconnection to Status and Respect	High Conscientiousness
Antifragility	See 3.4.1 paragraphs 3 and 4	Process	Reconnection to Meaningful Values	Low Neuroticism
Mastery	See 3.4.1 paragraph 5	Outcome	Reconnection to Meaningful Work	High Conscientiousness

2.4.1 Presence, Antifragility, and Mastery Badges

The Presence Badge tracks the degree to which the teacher was able to feel the presence of a student either through (1) attendance during live lectures, (2) submitting code to programming-type assessments or (3) in case the student is having a hard time understanding the exercises and cannot submit, presence may be secured through submitting summaries of the slides given for the week. The teacher prepares a weekly report of the students' presence and informs the student through a private profile. We associate this mechanics with the Identity layer of behavior change because giving students access to how their teacher perceives their presence while giving students a weekly chance to change how their presence is appraised or seen (identity) is connected to identity building. This act of identity building aligns with Hari's reconnection to status and respect concept, where it is explained that both are gained when a person is able to see the capability to produce consistent output (in this case, the output is weekly presence). This mechanic also aligns with High-Conscientiousness people as these personality types "enjoy participating in social rules that happen in a set time/interval".

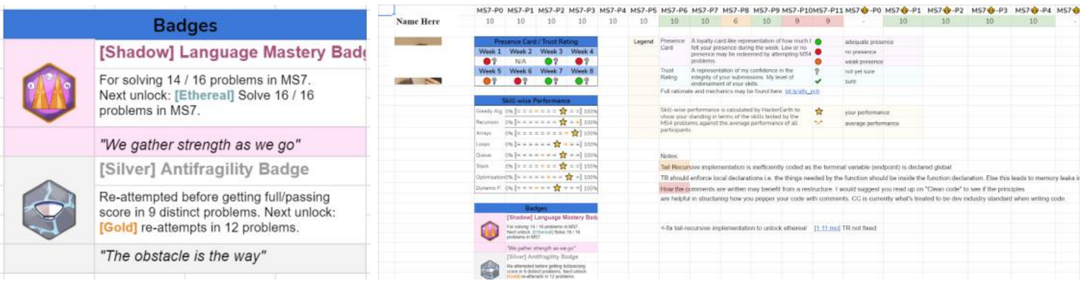
Antifragility is a concept recently that speaks of the ability of a person to not only recover from an unpleasant experience but become better for it (Taleb, 2012). In the context of academia, a student who is antifragile may be someone who would get a B+ on a third quiz after getting unsatisfactory marks on quizzes one and two. In the context of this paper, the antifragility badge is levelled up when a student is able to submit a correct solution/code to the online judge after submitting at least 1 incorrectly judged code submission. We associate this mechanics with the Process layer of behavior change as it directly encourages the (process) of revision. The ability to "bounce back from defeat or stressful events" is also a trademark of Low-Neuroticism individuals. We can also argue that the concept of anti-fragility is a proper subset of a meaningful value more commonly known as perseverance.

Finally, the Mastery badge tracks the number of competitive programming problems solved for the duration of the class. It reconnects the student to meaningful work as correctly answering such exercises validates the command of a relevant skill. It is designed for High-Conscientiousness individuals as "high attention to detail" (programming) and "focus on important tasks" (i.e. tasks that score points) are appealing to this personality type.

3. Methodology

3.1 Low-Fidelity Gamified Tracker

A low-fidelity prototype of a gamified tracker was implemented to test the design of the badge mechanics. In this prototype, each student has their own corresponding sheet that simulates a user profile. It has their name, photo, and the class activities with their scores. It also has a Presence Card -- an alternative way of checking attendance -- and a Trust Rating -- a mechanism for teachers to communicate with their students their perceived legitimacy of their submitted work (Casano & Agapito, 2021 in press). The student trackers also housed the badges earned by the student. The badge information includes the badge name, badge level (e.g. bronze, silver), and a description of how the badge was earned and what is needed to rank up the badge to the next level. See Figure 2 for a closer look of the badges earned by a student.



Badges

[Shadow] Language Mastery Badg

For solving 14 / 16 problems in MS7.
Next unlock: [Ethereal] Solve 16 / 16 problems in MS7.

"We gather strength as we go"

[Silver] Antifragility Badge

Re-attempted before getting full/passing score in 9 distinct problems. Next unlock: [Gold] re-attempts in 12 problems.

"The obstacle is the way"

Name Here	MS7-P0	MS7-P1	MS7-P2	MS7-P3	MS7-P4	MS7-P5	MS7-P6	MS7-P7	MS7-P8	MS7-P9	MS7-P10	MS7-P11	MS7-P0	MS7-P1	MS7-P2	MS7-P3	MS7-P4	MS7
	10	10	10	10	10	10	10	10	6	10	9	9	-	10	10	10	10	-

Presence Card / Trust Rating

Week 1	Week 2	Week 3	Week 4
🔴 ?	N/A	🟢 ?	🔴 ?
Week 5	**Week 6**	**Week 7**	**Week 8**
🟠 ?	🔴 ?	🟢 ?	🟢 ?

Skill-wise Performance

Greedy Alg	0% [= = = = = = = ☆ = =]	100%
Recursion	0% [= = = = = = = ☆ = =]	100%
Arrays	0% [= = = = = = = = = ☆]	100%
Loops	0% [= = = = = = ☆ = = =]	100%
Queue	0% [= = = = = = = ☆ = =]	100%
Stack	0% [= = = = = = = ☆ = =]	100%
Optimization	0% [= = = = = = = = ☆ =]	100%
Dynamic P.	0% [= = = = = = ☆ = = =]	100%

Badges

[Shadow] Language Mastery Badg

For solving 14 / 16 problems in MS7. Next unlock: [Ethereal] Solve 16 / 16 problems in MS7

"We gather strength as we go"

[Silver] Antifragility Badge

Re-attempted before getting full/passing score in 9 distinct problems. Next unlock: [Gold] re-attempts in 12 problems.

Legend

Presence Card	A loyalty-card-like representation of how much I felt your presence during the week. Low or no presence may be redeemed by attempting MS4 problems.	🟢	adequate presence
		🔴	no presence
		🟠	weak presence
Trust Rating	A representation of my confidence in the integrity of your submissions. My level of endorsement of your skills.	?	not yet sure
		✓	sure

Full rationale and mechanics may be found here: bit.ly/alb_pcfr

Skill-wise performance is calculated by HackerEarth to show your standing in terms of the skills tested by the MS4 problems against the average performance of all participants	☆	your performance
	^-^	average performance

Notes:

Tail Recursive implementation is inefficiently coded as the terminal variable (endpoint) is declared global

TR should enforce local declarations i.e. the things needed by the function should be inside the function declaration. Else this leads to memory leaks i

How the comments are written may benefit from a restructure. I would suggest you read up on "Clean code" to see if the principles are helpful in structuring how you pepper your code with comments. CC is currently what's treated to be dev industry standard when writing code.

<-fix tail-recursive implementation to unlock ethereal [1.11 mo] TR not fixed

Figure 2. Sample Badges and Gamified Low-Fidelity Tracker.

This prototype was manually updated by the teacher at the end of each school week. A student's "profile" (i.e., their corresponding sheet) was only shared with the profile's owner. They did not have access to the profiles of their classmates.

3.2 Testing

The low-fidelity gamified tracker was implemented in three (3) undergraduate classes in Ateneo de Manila University (ADMU), Manila, Philippines -- one (1) section of CSCI 20 - Introduction to Computing (section A) and two (2) sections of CSCI 30 - Data Structures and Algorithms (sections A & B). These were held during the first and second quarters (Q1 & Q2), respectively, of the school year 2020-2021. The quarterly setup was adopted by ADMU to help students cope with online learning during the pandemic to allow them to take fewer classes each quarter. Ideally, each quarter is 7-8 weeks long. However, due to schedule adjustments because of COVID-19 and other unforeseen circumstances such as typhoons, Q1 had only six (6) weeks while Q2 had eight (8). Both classes had programming activities and were handled by the same teacher.

The tracker was used in both classes to provide a more gameful method of tracking their class progress. It was also used as a tool to test the design of the different badge mechanics. At the end of their respective quarters, a short survey was conducted to collect the students' feedback about the use of the gamified tracker. It asked whether they were using/checking the tracker, their perceived utility of the badges, and their suggestions for improvement. It also sought their insights about the specific badges with the following items: (1) Would you say that the Language Mastery badge was a motivator for you to engage (look at, code a solution, make a submission) with the asynchronous online problem set? (2) Would you say that the Anti-fragility badge was a motivator to keep trying to code a solution towards full points? (3) Would you say that the Anti-fragility badge was helpful in letting you cope with (and/or bounce back from) an incorrect submission? (4) Would you say that the Presence badge was a motivator to engage with the learning material/or online problem set even when you felt you were stuck?

Nineteen (19) of the twenty-six (26) CSCI 20 students and nineteen (19) of the forty-four (44) CSCI 30 students responded to the survey. Results are presented in the next section.

4. Results and Discussion

To determine how the Xiphias badges affected student learning experiences (RQ2), we (1) attempted to look at any patterns in the badges earned by students with respect to their final letter grades; and (2) summarize the insights collected through the survey.

4.1 Badges Earned and Final Letter Grades

In this study, we looked at the badges earned by the exceptional students (letter grade of A) and those who received satisfactory grades of C or C+ in the two classes.

Nine (9) of the twenty-six (26) (35%) CSCI 20 students received a final letter grade of A, all of which earned Gold for Language Mastery. For CSCI 30, twenty-one (21) of the forty-four (44) (48%) students got an A and all of them received high Language Mastery (15 Ethereal, 5 Shadow, and 1 Gold). Additionally, most of the A-graded students earned the three badges. This shows that A students generally perform well in the programming exercises/assignments which were the primary basis for their grade. Also, a common thread among them is the ability to correct an incorrect submission. They take advantage of the opportunity to re-attempt failed work and such behavior allows them to gain mastery in the skills taught in these classes. Sustained presence in class either synchronously or asynchronously likewise characterizes these students.

The students who received a satisfactory grade of either C or C+ had patterns quite distinct from the A students. The one student who received a C+ in CSCI 20 got a Bronze Language Mastery Badge and did not receive a Presence Badge or an Antifragility Badge. For CSCI 30, four (4) students received a C and none of them received a Presence Badge. In terms of Language Mastery, three (3) of them got Silver while one (1) got Gold. Two (2) of these students did not earn the Antifragility Badge while one (1) got gold and one (1) got silver. There were two (2) students who got a C++ and both of them earned Bronze Presence Badges. One of them did not receive the Language Mastery and Antifragility Badges.

The other got Silver in Language Mastery and Bronze in Antifragility. Most of the students who received a satisfactory grade did not receive Presence badges but received Silver or Gold Mastery badges. These students were not present on most of the weeks but have attempted to answer the problems around the end of the quarter. Only a few among the receivers of C and INC received the antifragility badge, as there was only a little attempt from these students to improve their solutions to receive a more satisfactory grade. There were also four (4) students who received an INC in CSCI 20 and all of them did not receive presence and mastery badges. Overall, most of the students who received passing marks have received at least two badges.

These results are not in any way conclusive and generalizable. This is an initial attempt at examining the badges earned by students with respect to the letter grades they received. A more in-depth exploration of the data would be necessary to make more sound insights about these patterns.

4.2 Student Insights

A total of 38 students responded to the survey (19 from CSCI 20 and 19 from CSCI 30). Thirty-one (82%) indicated that they were checking their respective trackers. More than half of those who checked (21/31 or 68%) were motivated by the Language Mastery Badge. One student found it nice to receive it as a form of validation but was explicit in saying that it did not necessarily motivate them. Seventy-four percent (23/31) said the Antifragility Badge was a motivator to keep trying to code a solution towards full points and that it was helpful in letting them cope with and/or bounce back from an incorrect submission. The Presence Badge was a motivator for 25 of the 31 students (81%) in terms of driving them to engage with learning materials and/or problem sets even when they were feeling stuck.

Most of the respondents gave positive feedback on the badges as indicated by the numbers above. These results are further supported by students' responses when asked about their perceived utility of the badges. The students found them helpful and motivating. One student shared how "... it sorts of turns the task into a game with corresponding achievements and awards..." Another student "thought they were pretty useful, they motivated [them] to keep doing [their] best (and beyond) to get those badges (plus they looked cool so it motivated [them] even more to try and get them to look even cooler)." The students were also able to track their progress and helped them "understand where [they] stood with the class.". Some found the badges as a nice touch to validate them on their progress as well as assured them that they were on the right track. One student suggested that perhaps a badge list (similar to an achievement list in games) would further motivate and engage students as this allows them to see what they can work for.

Feedback such as these indicate the potential of the new Xiphias badge mechanics to provide students with a more gameful experience. However, whether the motivation is driven internally (i.e. based on the design framework discussed above) or externally (i.e. purely because they wanted to earn the badge) is something that will require further exploration and analysis.

5. Conclusion

This study is an attempt to design a gamification framework that is more deeply rooted in user motivations. We designed three new Xiphias Badges -- Presence, Mastery, and Antifragility -- based on the merging of salient features of James Clear's Behavior Change model (2016), Johann Hari's Lost Connections model (2018), and Jordan Peterson's Big Five model (2007). Designing the badges using a multidimensional approach is an attempt to cater to the multidimensionality of a user and hopes to be a more motivating gamification approach that taps into their internal motivation.

An initial prototype of the badges was implemented in undergraduate classes. Indicative patterns in the badges earned by students showed that A students are likely to receive the three badges with high levels in Mastery. These students were likewise antifragile in terms of choosing to re-attempt failed work to be able to get a better score. This behavior may imply that students who correct incorrect submissions are more likely to gain skill mastery. Satisfactory students (C or C+ student) did not have as much presence as the A students as depicted by the Presence Badge results. The results of a survey that sought their insights on the utility of the badges showed their potential to be motivating factors in the classroom. However, whether the motivation is intrinsic or extrinsic is something that needs to be further investigated. Findings from this study are not conclusive nor generalizable. This is an initial attempt to test the new badge designs. Nonetheless, the positive feedback from students reaffirms the potential of gamification as a tool to make learning more fun and motivating.

Acknowledgements

We thank the University Research Council of the Ateneo de Manila University for the grant entitled "*Xiphias: Using A Multi-Dimensional Approach Towards Creating Meaningful-Gamification-Based Badge Mechanics*" and the students who participated in our data gathering process.

References

Agapito, J. L., Martinez, J. C., & Casano, J. D. (2014, October). Xiphias: A competitive classroom control system to facilitate the gamification of academic evaluation of novice C++ programmers. *In Proceedings of International Symposium on Computing for Education, ISCE* (Vol. 14, pp. 9-15).

Bakkes, S., Tan, C. T., & Pisan, Y. (2012, July). Personalised gaming: a motivation and overview of literature. *In Proceedings of the 8th Australasian Conference on Interactive Entertainment: Playing the System* (p. 4). ACM.

Casano & Agapito (2021, in press). Towards the Design of an Adaptive Presence Card and Trust Rating System for Online Classes. *Adaptive Instructional Systems. Design and Evaluation*. Springer Nature Switzerland AG

Chou, Y. K. (2019). Actionable gamification: Beyond points, badges, and leaderboards. *Packt Publishing Ltd.*

Christy, K. R., & Fox, J. (2014). Leaderboards in a virtual classroom: A test of stereotype threat and social comparison explanations for women's math performance. *Computers & Education,* 78, 66-77.

Clear, J. (2018). Atomic habits: An easy & proven way to build good habits & break bad ones. *Penguin.*

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011, September). From game design elements to gamefulness: defining gamification. *In Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments* (pp. 9-15). ACM.

Garner, S., Haden, P., & Robins, A. (2005, January). My program is correct but it doesn't run: a preliminary investigation of novice programmers' problems. *In Proceedings of the 7th Australasian conference on Computing Education-Volume 42* (pp. 173-180). Australian Computer Society, Inc.

Hamari, J., Koivisto, J., & Sarsa, H. (2014, January). Does gamification work? --a literature review of empirical studies on gamification. In *2014 47th Hawaii international conference on system sciences (HICSS)* (pp. 3025-3034). IEEE.

Hari, J. (2019). Lost connections: Uncovering the real causes of depression-and the unexpected solutions. *Bloomsbury Publishing Plc*.

Ibáñez, M. B., Di-Serio, A., & Delgado-Kloos, C. (2014). Gamification for engaging computer science students in learning activities: A case study. *IEEE Transactions on learning technologies*, 7(3), 291-301.

John, R., & John, R. (2020). The Big Five Personality Traits and Academic Performance. *Journal of Law & Social Studies (JLSS)*, 2(1), 10-19.

Neve, P., Livingstone, D., Hunter, G., Edwards, N., & Alsop, G. (2014). More than just a game: Improving students' experience of learning programming through gamification. *Online*:< http://www. heacademy. ac. uk/system/files/comp-224-p. pdf>. Data dostępu, 13.

Peterson, J. B., DeYoung, C. G., & Quilty, L. C., (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of personality and social psychology*, 93(5), 880.

Sailer, M., & Homner, L. (2020). The gamification of learning: *A meta-analysis*.

Satow, L. (2021). Reliability and Validity of the Enhanced Big Five Personality Test (*B5T).*

Song, D., & Bonk, C. J. (2016). Motivational factors in self-directed informal learning from online learning resources. *Cogent Education*, 3(1), 1205838.

Taleb, N. N. (2012). Antifragile: Things that gain from disorder (Vol. 3*). Random House Incorporated*.

Tondello, G. F., Wehbe, R. R., Diamond, L., Busch, M., Marczewski, A., & Nacke, L. E. (2016, October). The gamification user types hexad scale. *In Proceedings of the 2016 annual symposium on computer-human interaction in play* (pp. 229-243).

Tondello, G. F., Mora, A., Marczewski, A., & Nacke, L. E. (2019). Empirical validation of the gamification user types hexad scale in English and Spanish. *International Journal of Human-Computer Studies*, 127, 95-111.

Watson, D., Nus, E., & Wu, K. D. (2019). Development and validation of the Faceted Inventory of the Five-Factor Model (FI-FFM). *Assessment*, 26(1), 17-44.

Tinkery: A Tinkerer's Nursery for Problem Solving with Lego Mindstorms

Ashutosh RAINA*, Sridhar IYER & Sahana MURTHY
IDP in Educational Technology, Indian Institute of Technology Bombay, Mumbai, India
*raina.ashu@iitb.ac.in

Abstract: Problem-solving, in most of the engineering design laboratories is still systematic by the book, lacking exploration, curiosity building, investigation and discovery. Even with the wide availability of tinkering kits designed in accordance with tinkerablilty, their ability to nurture tinkering is limited to few pre-built models and instruction manuals. In this paper we discuss building experiences of exploration and play as an operational understanding of tinkering. Guided by our understanding and tinkering practices we designed Tink-table a learning environment to nurture tinkering by solving engineering design problems in the domain of robotics. Tink-table uses a Lego Mindstorm kit, is supported by our XprSEv (read as expressive) pedagogy, has a set of problems and a mentor. This paper presents the design of Tink-table along with a study design to understand its role in nurturing tinkering. We present observations of a preliminary study that align with our understanding of tinkering.

Keywords: Tinkering, problem solving, lego mindstorms, robotics, learning environment

1. Introduction

The tinkering movement has gained tremendous momentum and one of its significant indicators in India was the establishment of Atal Tinker labs by The NITI Aayog Govt. Of India under the Atal innovation mission (AIM) of 2016. Under AIM 7200+ schools have Atal Tike labs established. Tinkering provides learners with a playful curious inductive perspective to problem and solution along with the deductive approach. Tinkering, not limited to engineering design, enables a learner with the skill of approaching the unknown and being able to explore and gain with the experience (Dym et al., 2005). The current generation has been provided with a variety and access to a lot of inclusive tinkerable tools and materials, some of which have been designed based on research to ensure tinerability and support tinkering ability. Based on our experience of using such kits we realize the support for nurturing tinkering abilities is limited to the pre-built models and instruction manuals which limits their potential as tools for the joy and love of tinkering and learning (Mitchel Resnick, 2017). The other challenge is that the formal setting for problem-solving, especially in the laboratories of engineering design, is very systematic and by the book, which does not encourage exploration, curiosity building, and the need for investigation and discovery (Atman & Bursic, 1996).

The broad objective of our research is to design a pedagogical strategy with a learning environment to nurture an attitude of tinkering for solving engineering design problems as an alternate approach to problem-solving. Also, to facilitate mentors in nurturing a tinkering attitude among problem solvers and learners. In order to reach the goals, we attempt to identify operational nuances of tinkering and use them to develop a learning solution for nurturing tinkering. In this paper we discuss our operational understanding of tinkering use it to design our learning environment called Tink-table and its components which are the XpeSEv pedagogy, building resources, problem statements and the mentor. We also present our observations from a single participant preliminary study.

2. Literature

Tinkering has been expressed in forms of various characteristics (Resnick & Rosenbaum, 2013; Stager & Martinez, 2013), narratives (Dougherthy, Honey, & Kanter, 2013) and as ways of life (Louridas, 1999). Tinkering has been considered as a novice and expert practice which sets it apart from most of the classroom practices (Danielak et al., 2014). It does not make tinkering better or worse but it does make it an authentic professional practice (Berland et al. , 2013). Tinkering is also associated with "jugaad" and "bricolage". From bricolage comes an attitude of building experiences from the immediate sensory perceptions by exploration and experimentation (Louridas, 1999) and from jugaad the ideas of starting with a quick fix (Radjou et al., 2012). In contrast, deliberate sensemaking aims at conceptual understanding but in tinkering producing an outcome is a primary goal. It could drive deliberate sensemaking, only in service of the outcome (Quan & Gupta, 2020).

Research in engineering design has recognized that iteration and experimentation supports generation of knowledge and designs refinements (Dym et al., 2005). With rapid prototyping, the generation of manipulable artifacts from the initial design ideas to refine the design (Berland et al., 2013), overlaps with tinkering as it is improvisational and iterative toward the design goals (Baker et al., 2008). In design, the goal is to produce an artifact or solution while the terms of success are not well-defined and multiple solutions approaches are possible. The engineering design process also uses multiple approaches (Dym et al., 2005), while solving a complex problem (Jonassen, 2000), tinkering is one such practice of the engineering design process. Tinkering, like all engineering activities, is a situated phenomenon (Johri et al., 2011). Tinkering emerges within interactions between students and their in-the-moment goals and is sustained by feedback from the social and material environment. Tinkering as per our understanding, is a disposition based on inquiry into the problem and solution space further mapping it to one's associated knowledge while solving it. Tinkering is an iterative experience-driven approach which is full of short and long cycles of quick experimentation which results in observations leading to a different or new understanding. Through these short and long cycles emerges an improved understanding with which the solution continues the evolution.

3. When We Tinker

Many discussions associate tinkering with playful explorations at its core (Mitchel Resnick, 2017). Two distinct interactions, emerge as important; *exploration* - which is asking the question "*what things do?*" and *play* which is "*what can I do with them?*" (Zubrowski, n.d.). These activities allow learners to build repositories of experiences. An experience is an interaction and/or an investigation that gives rise to observations developing some understanding. The learners later draw from these experiences when faced with challenges (Dyasi, n.d.). This aligns to classic bricolage literature where (Louridas, 1999) for a tinkerer there is a repository of experience-based understanding that comes from asking the questions "*what can something do*?" (Exploration) and "*what do I want to do with it*?" (Play). So, problem solving for a tinkerer is having interactions with problem and solution space which consist of a lot of explore and play to build some understanding about them. Then start with a quick fix and continue to evolve as understanding develops. Hence, Our understanding of a tinkerers way to problem solving is1) Exploration of resources (solution space) with the question "*What can the resources do?*" and problems (problem space) with the question "*What do I want to do for the problem?*" and 2) Play with the understanding of what things "*can do*" and the things I "*want to do*" to create a mapping between them by answering the question "*What should the resource do for me?*". The answers either lead to modification of ones understanding of the resources or the way they perceive the problem. Through the evolution of these understanding of the resources and the problem the solution evolves. With this as an operational understanding as our theoretical framework we strive to nurturing an attitude of exploration and play, or for someone who already does it, make it explicit.

4. Nurturing Tinkering

Based on the operational understanding of tinkering, literature and a few pervious explorations (Raina et al., 2018, 2019) we have designed Tink-table our learning environment. Tink-Table is built on four aspects namely The XprSEv (read as expressive) pedagogy, A set of problems from the domain of robotics, building resources which in our case is Lego Mindstorms, and a mentor.

4.1 The XpreSEv Pedagogy

The pedagogue is based on our previous explorations with tinkering and derives from the models of tinkering (Mitchel Resnick, 2017). The elements pertaining to Tinkerability and supporting tinkering ability have also been considered (Mitchel Resnick, 2017). The pedagogy is primarily based on three objectives namely explore, solve and evolve. With explore, the learners focus on learning to explore the problem and solution space. In Tink-table this objective is operationalised into phase 1 where learners start with small challenges situated in context robotics, which requires them to interact with the physical space using the components of the robotics kits to understand their affordances to solve these challenges. With solve, the learners focus on play by mapping their understanding of the resources to what they would want to do to solve the problem. In Tink-table this objective is operationalised in phase 2 in addition to previous objectives where the learners use their understanding of the robotics kit and try to map it to a way, they would want to approach the given problem. If a mapping is not achieved the learner either updates their understanding of the use of the resources from the robotics kit or works on a different way to solve the problem based on what resources are available. With evolve, the learners focus on evolving their solution in terms of structure or function. In Tink-table it is operationalized in phase 3 along with phase 1 and 2 where the learners are asked to frame a new problem in robotics or think of a way of solving an emergent problem in the solution of the previous given problem.

The challenges given in phases 1 are candidate sub problems for the problem given in phase two and three. This is done adhering to on progressive formalization. The Objectives determine the focus of problems designed for each phase and the activities to be performed by the mentor. A summary of the same has been present in Table 1.

Table 1. *Operationalization of the XprSEv Pedagogy into Three Phases.*

Phases	Objectives	Activity Focus	Learner Goals
1	Explore	Explore resource affordance and use them to solve candidate problems	Understand and use resources as per affordances and find the affordances required for their solution approach.
2	Solve	Solve problem by finding solutions for sub problems with the given resources.	Divide into subproblems & identify affordances. Use resources based affordances OR solve sub problems as per affordances of the resources available.
3	Evolve	Improve solution or solve emergent challenges by refining sub problems while reflecting on interaction.	Improve the solution by exploring alternate resources for sub problems and playing with their affordances.

4.2 The Problems

The problems are designed to progressively complicate yet allow the learners to connect their understanding from one to the other allowing progressive formalisation with the resources and the materials available. The problems are based on the Lego kit. Learners are initially given challenges that nudges them to explore the affordances of the components available in the kit. The challenges require them to use a particular affordance of the resource offered in the kit in a way they are able to explore the possibilities of what the resources can do. The next problem is based on the challenges given initially. Finally, the learners try to refine the solution of the problem or attempt a problem is of their choosing.

4.3 The Mentor

The role of the mentor is of a non-contributing participant. The mentor has to be very well versed with the entire problems by solving them and exploring variations. It helps mentors to empathizes with the learners and scaffold them towards exploration and play in terms of can do’s and want to do’s by posing questions. The mentor's approach is shared by talking about how and why they thought or suggested certain possibilities. It is recommended to give multiple possible approaches and leave the decision of trajectory to learner. The mentors could intervene if they encounter learners in discomfort and scaffold them towards flow. The mentor foresee a lot of challenges in the learner's trajectory, but not interrupt the learner unless learners ask for help. If the mentor does foresee failure, should allow the learner to observe failure but question later. It is a crucial step to develop an understanding about the resources and the solution approach. One preferred approach is to re-articulate the questions posed by the learners which brings clarity that the learner and the mentor are on the same page and also acts as a trigger for reflection for the learner. Mentor can bring the learner to a point where lot of possibilities exist without explicitly mentioning them. Then get the learner to think about possibilities or give suggestions by using the resources as an aid to communicate.

4.4 The Resources

As all our problems are based on robotics, we have used a Lego Mindstorms kit as a building resource for our studies with Tink-table. This kit is designed with the principles of Tinkerability and also supports tinkering ability. Moreover, by limiting the type of building resources i.e., the kit, we as researchers have been able to develop a thorough understanding of the resources and their affordances and this will help us interpret the interactions of the learners with these resources. Additionally, some semi-built models built of Lego are kept as scaffolds for the learners to understand how the components fit together. The learners also have a system that allows the learners to program Lego or even browse online resources.

5. Study Design and Preliminary Observations

The study was designed as a workshop for learners to experience tinkering by means of solving a set of problems over three days three hours each day using the Lego Mindstorm kit. The pedagogical design was operationalised into three phases with objectives as discussed in the above section. Due to the restrictions of COVID-19 and closure of the lab facility a preliminary study was conducted in an online mode where the resources were sent to the participant and set up in their location. The observer and the mentor joined via online virtual conferencing. The participant's and the mentor's video and audio were always on where-as the observer's audio and video was always off. The participant was from the 9th standard. The participant had not worked with Lego Mindstorm or any such robotics kit previously. The participant had limited exposure towards programming with a few hello world python programs. When the participant was asked how she felt about working with the Lego kit she said she wanted to build robots, but she was daunted as she did not know programming and had worked with such a kit.

We conducted the preliminary study over four days with a gap of 4 days between each study day. In Phase 1 (study day 1) the participant were to 1) use the Lego brick with any sensor deemed fit to a) measure the area and then the volume of the room b) sense primary colors and identify them on the Lego brick screen; and 2) Build a four-wheel robot that can move forwards or backward. The focus of challenge 1 was to ensure the learner experiments with Lego Mindstorms's Ultrasonic Sensor, IR / Color sensor able to understand the affordances of distance measurement, color detection, proximity detection. The other challenge focused on the building resources like beams, pegs (3 types), wheels, frames and angle joints to understand their affordances in different forms of construction and connection. The mentor was to observe the learner's trajectory towards the solution and provide prompts and direct operational information about components. It was orchestrated by the mentor using questions to trigger reflection. E.g., "*Why did you drop the previous idea ?*" encourage play to look for feasibility E.g. "*Why don't you try it out*" and direct them towards scaffolds E.g. "*See how it has been used there*?" The mentor in this phase was more of a non-contributing co-learner. In Phase 2, (study day 2 & 3) The problem to be solved by the participant was to make a bot which could 1) move front, back and turn. 2) navigate a marked grid of tiles by using control buttons on the brick and later with a cabled remote also made with the Lego Kit and 3) navigate the maze autonomously. Examples of a few objectives were to choose between a 2 motor or 4 motor design, determine the algorithm for the motors to enable turning right, left and back; a structure that allows swift turning of the bot, building the bot in a way it could be driven by push buttons built remote. For autonomous mode, estimating the distance to be traveled and find equivalent rotations to be programed and estimating the rotation direction and angle for each motor to make a left and right turn. The mentor was to observe the learner's actions on the components being used and trigger them to experiment and play by guiding them to reflect on actions of previous phase. They also did rain checks ensure flow and if stuck, for e.g., in deciding wheel configuration, they provided alternates as analogies like 3-wheel vs 4-wheel design. Mentors also directed the learner towards digital manuals for ideas and exploring functions of the components. The mentor's role in this phase was of a non-contributing co-creator. For phase 3 (study day 4) the participant was asked to build something of personal interest, and she choose to build an automatic dog feeder. The mentor in this phase took a step back and just observed the participant. Mentors choose to provide operational information and had some conversations to develop an understanding of what the learner's idea is and what they intend to do. In our case the mentor who was also the researcher who took an open interview enquiring about how she started by imagining the dog feeder in terms of Lego components, how the solution evolved over time quoting instances noted down during the observation and ending the interview by giving her a scenario and asking her how she would build the solution in terms of the Lego components.

6. Preliminary Study and Findings

The broad question of the preliminary study was, does working with Tink-table nurture behavior synonymous to operational definition of tinkering, i.e., do learners build an understanding of affordances of Lego components? Do learners think of ideas in terms of the Lego Kit components ? Does the learner trajectory involve evolution of their ideas as understanding develops? The entire online sessions were recorded in audio as well as video focusing on the interactions of the learners with the robotics kit and the mentor. The observer took notes on similar interactions. The open interview was recorded in audio and video. The recordings were used to identify episodes of interactions between the learner and the Lego Mindstorm kit and determine if the current interactions represented signs of exploration. Based on the classification of the episodes and their implications the observations were classified into a) repository of resource capabilities, b) ability to express ideas as artifacts with resources, and c) evidence of evolution from a quick fix.

The observations suggest that the participant showed signs of acquiring knowledge about the affordances of the materials in a number of instances. Though the participant preferred structural solutions, by the last day she was using and had understood the difference in the affordance of the two different motors. The participant demonstrated that she could connect the beams in different configurations by using different types of pegs to achieve different structures. Similar instances made it evident that the participant by the end of fourth day had developed an understanding of how various pieces of the kit could be used and was able to use them in various ways to achieve her objectives. During conversations with the participant internalization of form and function of Lego pieces was evident. She was also able to translate her idea of moving the motors slowly which she described in terms of angle of rotations per min and then materialized that same by controlling power that was moving the motor. Instances show that the participant was consistently thinking of improving the solution at hand which became evident as the workshop progressed. She was initially struggling with challenges and the fear of not having experience with robotics seemed to take over but after solving the first challenge of measuring the volume of the room got her the thought process going. She said she liked this approach of asking questions and trying to figure out things on her own. It was certainly more fun than someone giving instruction.

7. Conclusion

The initial observations suggest with Tink-table could lead to internalization of the Lego kit pieces and participant's ability to think in terms of the pieces when asked to build an artifact. Additionally, the participant has shown evidence of being able to externalize ideas as products made with Lego kit pieces. Progressive formalization with Lego and the way challenges have been designed will allow participants to build interest and experiment to develop an understanding of how they work and what they can do. The elements of Tink-table seem to support tinkering-based approach in addressing problems and has shown evidence that such an approach of mapping the “can do’s” and “want to doo’s” can be inculcated via Tink-Table. Additionally, a change was observed in participant's confidence and acceptance to this playful style of problem-solving grounding her interest into domain of robotics which initially felt daunting. This study was limited to a single participant due to the covid restrictions, but the insights are promising of nurturing a tinkering approach to problem solving with Tink-table. As facilities re-open we aim at conducting more in lab studies to also understanding the underlying processes that provide strength to a tinkerer's approach.

References

Atman, C. J., & Bursic, K. M. (1996). Teaching engineering design: Can reading a textbook make a difference? *Research in Engineering Design - Theory, Applications, and Concurrent Engineering*, *8*(4), 240–250. https://doi.org/10.1007/BF01597230

Baker, D., Krause, S., & Purzer, S. Y. (2008). Developing an instrument to measure tinkering and technical self-efficacy in engineering. In *ASEE Annual Conference and Exposition, Conference Proceedings*. American Society for Engineering Education. https://doi.org/10.18260/1-2--3413

Berland, M., Martin, T., Benton, T., Petrick Smith, C., & Davis, D. (2013). Using Learning Analytics to Understand the Learning Pathways of Novice Programmers. *Journal of the Learning Sciences*, *22*(4), 564–599. https://doi.org/10.1080/10508406.2013.836655

Danielak, B. A., Gupta, A., & Elby, A. (2014). Marginalized Identities of Sense-Makers: Reframing Engineering Student Retention. *Journal of Engineering Education*, *103*(1), 8–44. https://doi.org/10.1002/jee.20035

Dougherthy, D., Honey, M., & Kanter, D. E. (2013). *Design, Make, Play: Growing the Next Generation of STEM Innovators*. (M. Honey & D. E. Kanter, Eds.). Routledge.

Dyasi, H. (n.d.). Pedagogical Perspective: Hubert Dyasi. Retrieved May 31, 2021, from https://www.coursera.org/learn/tinkering-motion-mechanisms/lecture/qDZXc/pedagogical-perspective-hubert-dyasi

Dym, C. L., Agogino, A. M., Eris, O., Frey, D. D., & Leifer, L. J. (2005). Engineering design thinking, teaching, and learning. In *Journal of Engineering Education* (Vol. 94, pp. 103–120). Wiley-Blackwell Publishing Ltd. https://doi.org/10.1002/j.2168-9830.2005.tb00832.x

Johri, A., Olds, B. M., Esmonde, I., Madhavan, K., Roth, W. M., Schwartz, D. L., … Tabak, I. (2011). Situated engineering learning: Bridging engineering education research and the learning sciences. *Journal of Engineering Education*, *100*(1), 151–185. https://doi.org/10.1002/j.2168-9830.2011.tb00007.x

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development*, *48*(4), 63–85. https://doi.org/10.1007/BF02300500

Louridas, P. (1999). Design as bricolage: Anthropology meets design thinking. *Design Studies*, *20*(6), 517–535. https://doi.org/10.1016/s0142-694x(98)00044-1

Mitchel Resnick. (2017). *Lifelong Kindergarten: Cultivating Creativity Through Projects, Passion ... - Mitchel Resnick, Ken Robinson - Google Books*. The MIT Press.

Quan, G. M., & Gupta, A. (2020). Tensions in the productivity of design task tinkering. *Journal of Engineering Education*, *109*(1), 88–106. https://doi.org/10.1002/jee.20303

Radjou, N., Prabhu, J., & Ahuja, S. (2012). *Jugaad Innovation: Think Frugal, Be Flexible, Generate Breakthrough Growth - Navi Radjou, Jaideep Prabhu, Simone Ahuja*. (K. Roberts, Ed.). Jhon Wiley & Sons.

Raina, A., Murthy, S., & Iyer, S. (2018). "Help me build": Making as an enabler for problem solving in engineering design. In *Proceedings - IEEE 18th International Conference on Advanced Learning Technologies, ICALT 2018* (pp. 455–457). IEEE. https://doi.org/10.1109/ICALT.2018.00113

Raina, A., Murthy, S., & Iyer, S. (2019). Designing TinkMate: A Seamless Tinkering Companion for Engineering Design Kits. In *Proceedings - IEEE 10th International Conference on Technology for Education, T4E 2019* (pp. 9–14). IEEE. https://doi.org/10.1109/T4E.2019.00-58

Resnick, M., & Rosenbaum, E. (2013). Designing for tinkerability. In M. Honey & D. Kanter (Eds.), *Design, make, play : growing the next generation of STEM innovators*. New York, NY : Routledge.

Stager, G. S., & Martinez, S. (2013). *Invent To Learn: Making, Tinkering, and Engineering in the Classroom*. *undefined*. Constructing Modern Knowledge Press.

Zubrowski, B. (n.d.). Inspiration: Bernie Zubrowski | Coursera. Retrieved May 31, 2021, from https://www.coursera.org/learn/tinkering-motion-mechanisms/lecture/6q5Le/inspiration-bernie-zubrowski

Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Birds of Paradise: A Game on Urban Bird Biodiversity Conservation

James Matthew L. CUARTERO, Patricia Vianne C. LEE[*] & Jamielyn Mae C. VILLANUEVA
Ateneo Laboratory for the Learning Sciences, Ateneo de Manila University, Philippines
*patricia.lee@obf.ateneo.edu

Abstract: Birds of Paradise is a mobile game that aims to raise awareness of urban bird biodiversity. Due to a disconnect with nature, humans are becoming more indifferent to biodiversity, including urban bird biodiversity. This disconnect leads humans to take part in activities that harm biodiversity. The researchers believe that through educating the players on different bird species and the importance of green spaces, the game can aid environmental literacy specifically on urban bird biodiversity conservation. The 2D collecting game Birds of Paradise is developed for Android mobile devices where players get to build their own green space, play a minigame, and complete missions, all of which will help the player learn more about urban birds.

Keywords: Mobile Game, anchored instruction, game-based learning, urban bird biodiversity, bird biodiversity conservation

1. Introduction

The Philippines has a rich avifaunal diversity and houses more than 600 different resident and migrating bird species with new bird species still being discovered to this day. One third of these bird species are endemic to the Philippines. The bird population of the Philippines makes up around 6% of the world's total bird population; however, one third of the bird species that can be found in the Philippines are threatened, and these numbers are on the rise. This negative trend on birds is caused by several pressures, the majority being human activities such as deforestation, mishandling, and pollution (Panopio & Pajaro, 2014). Due to a disconnect with nature, humans are becoming indifferent to biodiversity and this disconnect leads humans to take part in human activities that harm the bird population (Miller, 2005).

One way to alleviate this disconnect is to spread environmental literacy, through the media or academe (Bickford et al., 2012). Hence, the researchers aim to develop a game that informs the players on how to help conserve birds in the urban landscape. By creating a game that is both informative and fun, the researchers believe that this can help spread environmental literacy specifically on urban bird biodiversity conservation.

Birds play an important role in the ecosystem. They are responsible for pest control, pollination, and other key activities that bring balance to the ecosystem (CGTN America, 2015). Due to this, it is important to protect bird species from various threats. A key challenge in wildlife conservation is the undervaluation of species (Belgrado, 2020). Through this study, the group aims to spread awareness on different bird species and the importance of green spaces. By educating the player on the aforementioned topics, the game aims to encourage players to value bird conservation.

2. Review of Related Literature

The goal of this research is to develop a game that can be used to educate and raise awareness on urban bird biodiversity conservation. In this chapter, the researchers explore how games can be used to educate using concepts such as Game-based Learning (GBL) and Anchored Instruction.

GBL is a type of gameplay with a defined learning outcome. GBLs are designed to balance gameplay with learning to help the player retain information from the game and apply it to the real world (EdTechReview, 2013). Games implementing the GBL technological paradigm have been shown to have positive effects on a student's motivation and engagement. Motivation pertains to the student's desire for learning and engagement is about the attention a student puts into whatever he/she is doing. Results of the research conducted by Serrano (2019) on students in elementary to high school level show that students have given positive feedback regarding GBL. Moreover, for a game to produce a positive effect on learning and engagement it is recommended by Schifter (2013) to have (1) Goal Orientedness, (2) Meaningful Interactions, (3) Engaging Narrative, (4) “Hero” Player, and (5) Eye-catching Visuals. Hence, to ensure that the game will be engaging and have a positive effect on learning, the researchers will apply the aforementioned characteristics.

Anchored Instruction focuses on using a type of media material to anchor the user to learning by solving complex, realistic problems (InstructionalDesign.org, 2018). The anchor in anchored instruction is the “scenario or situation given to learners that sets the stage or provides the context for use of learners’ knowledge or skills” (Glazer, 2014). Kariuki and Duran (2004) conducted research on anchored instruction wherein they used it to teach preservice teachers to integrate technology in the curriculum. The teachers used the educational computing class to record their experiences in the development class. The curriculum development class theme was used as the “anchor” for the educational computing class. The result of the research showed positive results in the preservice teachers’ learning. Given the effectiveness of GBL, the group will develop a game about urban bird biodiversity using the two principles of anchored instruction. The game will serve as the “anchor” for urban bird biodiversity learning. It will simulate real-life scenarios on creating green spaces and make the player identify the relationship of the birds to certain amenities. The game itself is interactive and will allow the player to explore the topic as long as it is within the scope.

3. Methodology

Since the main goal of the game is to inform by spreading awareness on urban bird biodiversity in the National Capital Region (NCR), Philippines, the group consulted subject matter experts who are knowledgeable on bird diversity and green spaces. This helped the researchers ideate and create a game that is as accurate as possible.

Birds of Paradise is a 2D bird collecting game, developed in Unity 2019.4.13f1 for Android devices, that aims to spread awareness on urban bird biodiversity. The player will be tasked to place amenities in their green space to attract birds. The main objective of the player will be to complete the bird catalog through discovering new birds. To do so, the player must place amenities in his/her green space as specific amenities attract specific birds. The game will also feature a mission system which encourages game progression. Asides from the main objective, the game will feature a minigame to test the player on their knowledge of the birds.

In order for the game to be engaging, it will be designed to cater to the behaviorism learning aspect of GBL. Since behaviorism is about learning through simulation and reinforcement, the game will have elements to represent this in specific features of the game. The three aspects of the game that would be modeled for a behaviorism approach would be as follows:

- Game Rules - The rules of the game are straightforward, no hidden mechanics or rules that need to be read in-between the lines.
- Gameplay - The game will implement a mission system to stimulate the player into achieving goals.
- Game Narratives - The game narratives would not be too information heavy. The player will learn more through the minigames and core gameplay.

In order to test the effectiveness of the game, the researcher conducted 2 rounds of user testing with 5 adult (18 years old and above) testers per round. The testing was remote and unmoderated where testers were given two questionnaires (pre-test and post-test) to answer. The questionnaires contain 4 main sections, 2 of which are similar in both the pre-test and post-test questionnaire. The two similar sections in both pre-test and post-test questionnaires are the learning effect section, which aims to measure the knowledge of the tester on urban bird biodiversity and green spaces, and the attitude effect section, which aims to measure the attitude of the tester towards urban bird biodiversity conservation.

The tester was tested before and after playing the game then the results were compared to see whether or not the game has a positive, negative, or no effect on the tester.

4. Conclusion

After conducting the testing, the researchers conclude that GBL can help aid the disconnect between humans and nature. The participants were able to retain information about the different bird species present in NCR and were also more familiarized with the concept of green spaces. Furthermore, the participants reported an overall positive game experience. Participants were able to immerse and challenge themselves with the game. Most importantly, participant's attitude towards urban bird biodiversity conservation improved. Participants were able to become more aware of the issues within urban bird conservation. The game that was developed by the researchers did not only focus on informing the player about urban bird biodiversity and green spaces but they also made sure that the game is both eye-catching and easy to use so that the player could focus on the learning aspect of the game.

Acknowledgements

We would like to express our gratitude to our advisers, Dr. Ma. Mercedes T. Rodrigo and Mr. Walfrido David A. Diy, our subject experts, Ms. Abigail Marie T. Favis and Ms. Maria Katrina C. Constantino, our family, and friends for their support throughout this research.

References

Belgrado, B. (2020, March 27). Combatting trade in illegal wildlife through awareness raising and a better understanding of consumer behaviour. Retrieved July 19, 2020, from https://www.niras.com/news/the-fight-against-illegal-wildlife-trade-in-the-philippines

Bickford, D., Kudavidanage, E. P., Campos-Arceiz, A., Qie, L., & Posa, M. C. (2012). Science communication for biodiversity conservation. Biological Conservation, 151(1), 74-76. doi:https://doi.org/10.1016/j.biocon.2011.12.016

EdTechReview. (2013, April 23). What is GBL (Game-Based Learning)? Retrieved July 20, 2020, from https://edtechreview.in/dictionary/298-what-is-game-based-learning#:~:text=Game%20based%20learning%20(GBL)%20is,that%20has%20defined%20learning%20outcomes.&text=Game%20based%20learning%20describes%20an,learning%20context%20designed%20by%20teachers

Glazer, E. (2014, October 18). Problem Based Instruction. Retrieved July 23, 2020, from http://epltt.coe.uga.edu/index.php?title=Problem_Based_Instruction

InstructionalDesign.org. (2018, November 30). Anchored instruction (John Bransford). Retrieved July 22, 2020, from https://www.instructionaldesign.org/theories/anchored-instruction/

Kariuki, M., & Duran, M. (2004). Using Anchored Instruction to Teach Preservice Teachers to Integrate Technology in the Curriculum. JI. of Technology and Teacher Education, 12(3), 431-445. Retrieved July 23, 2020, from http://www-personal.umd.umich.edu/~mduran/JTATE123431.pdf

Miller, J. R. (2005). Biodiversity conservation and the extinction of experience. Trends in Ecology &; Evolution, 20(8), 430-434. doi:10.1016/j.tree.2005.05.013

Panopio, J. B., & Pajaro, M. (2014). The State of Philippine Birds. Quezon City, Philippines: Haribon Foundation for the Conservation of Natural Resources. Retrieved from https://www.researchgate.net/publication/319624547_The_State_of_Philippine_Birds

Schifter, C. C. (2013). Games in learning, design, and motivation. In M. Murphy, S. Redding, & J. Twyman (Eds.), Handbook on innovations in learning (pp. 1–16). Philadelphia, PA: Center on Innovations in Learning, Temple University; Charlotte, NC: Information Age Publishing. Retrieved from http://www.centeril.org/

Serrano, K. (2019). The Effect of Digital Game-Based Learning on Student Learning: A Literature Review (Master's thesis, University of Northern Iowa, 2019). Graduate Research Paper. Retrieved July 23, 2020, from https://scholarworks.uni.edu/grp/943

Comparison of English Comprehension among Students from Different Backgrounds using a Narrative-centered Digital Game

May Marie P. TALANDRON-FELIPE[ab*], Kent Levi A. BONIFACIO[c], Gladys S. AYUNAR[c] & Ma. Mercedes T. RODRIGO[a]
[a]*Ateneo Laboratory for the Learning Sciences, Ateneo de Manila University, Philippines*
[b]*University of Science and Technology of Southern Philippines, Philippines*
[c]*Central Mindanao University, Philippines*
*maymarie.talandron-felipe@ustp.edu.ph

Abstract: This paper reports the continuation of the field testing of a narrative-centered digital game for English comprehension called *Learning Likha: Rangers to the Rescue* (LLRR) with a two-fold goal: first, identify the differences in terms of usage, attitudes towards, and perceptions of the English language between students from southern Philippines and the National Capital Region, and second, to determine how the LLRR in-game performance, post-test comprehension scores, engagement, and motivation of students differ between the groups . The participants who are grade school students from a province in southern Philippines answered questionnaires about their attitude towards and perception of English, played LLRR, answered the English comprehension post-test, and assessed their engagement and motivation using the adapted game-based learning engagement (GBLE) and intrinsic motivation inventory (IMI) questionnaires, respectively. Responses and interaction logs were compared to the data collected from NCR. Findings showed no significant difference between the groups in terms of the usage of English whether at home or with friends. However, NCR-based students were more receptive in terms of their perception and attitude towards the language, had better LLRR in-game performance, and obtained higher English comprehension post-test ratings. These findings are consistent with the results of the Programme for International Students Assessment (PISA) in 2018 where students from the southern regions have lower English reading literacy compared to those from NCR. In terms of GBLE and IMI responses, the gap is consistent as self-reports of participants from the south indicated lower behavior and emotion engagement, enjoyment, effort exerted, and perceived competence while playing LLRR.

Keywords: English comprehension, Philippines, game-based learning, mobile games, mobile-assisted language learning, narrative-centered digital game

1. Introduction

The Philippines is considered as one of the largest English-speaking countries in the world with about two-thirds of the population capable of speaking in the language. The country recognizes that mastery of English is beneficial in preparing globally competitive individuals as it is regarded as the world's lingua franca and the primary medium of communications for various global industries (Shobikah, 2017). The Philippines' English proficiency has led to the nation's leadership as a business process outsourcing (BPO) industry destination (Mariñas, 2021). The BPO industry in the Philippines is valued at US$23 billion, providing 1.15 million jobs (Lema, 2017). Further, the prevalence of English in higher education in the Philippines has attracted students from other non-English speaking countries like China and South Korea to study in the Philippines for undergraduate and postgraduate degrees (Romero, 2018).

However, in the recent Education First English Proficiency Index (2020), a noticeable decline has been observed as the Philippines dropped to 27th in the overall ranking out of 100 countries compared to being the 20th out of 100 in 2019 and 14th out of 88 in 2018 (EF EPI, 2020). Similarly, in its first year of participation in the Programme for International Students Assessment (PISA) in 2018, the Philippines ranked last among the 79 nations in reading proficiency. Only 19% of the students attained at least Level 2 proficiency in reading which means they can identify the main idea in a text of moderate length, find information based on explicit, though sometimes complex criteria, and can reflect on the purpose and form of texts when explicitly directed to do so. Unfortunately, almost no student (only 0.05%) attained level 5 in the PISA reading test which expects the student to comprehend lengthy texts, and to infer which information in the text is relevant even though the information of interest may be easily overlooked. This is relatively very low compared to15 other countries from the Organization for Economic Co-operation and Development (OECD) that have more than 10% of students who attained levels 5 and 6 (OECD, 2019).

The report indicated that 94% of the students in the Philippines who participated in the test do not speak the test language (i.e. English) at home. This led the test proponents to ask whether the choice of language affected the test performance. They also noted that expenditure per student in the Philippines was the lowest amongst all PISA-participating countries and 90% lower than the OECD average. Socio-economically advantaged students outperformed disadvantaged students in reading by 88 score points and students from private schools performed better than those from public schools by 62 points. Geographically, the National Capital Region's (NCR) mean score of 372 is 20 to 80 points higher than all other regions in the country while only the Central Mindanao Region in southern Philippines (Region 12) obtained a mean score lower than 300 (Philippine Department of Education, 2019). Given these findings, the government challenges the academe to review current curricula to improve English education and the engagement of all stakeholders particularly to support the learning of students from disadvantaged backgrounds (Romero, 2018).

The Ateneo Laboratory for the Learning Sciences with a grant from the Philippines' Commission on Higher Education and the British Council had undertaken a research project that developed various mobile applications to address the need for additional learning materials that could help students with English language learning (Ocumpaugh, Rodrigo, Porayska-Pomsta, Olatunji, & Luckin, 2018). Among the applications developed is a narrative-centered digital game for English comprehensions called *Learning Likha: Rangers to the Rescue* (LLRR). It exercises the skill of attention to details and understanding of instructions. LLRR has been initially tested on students from the NCR and findings show that the participants thought the game is interesting and fun, they were motivated to understand the content and complete the game tasks, and the comprehension scores were generally good (Agapito et al., 2020). Considering the findings in the PISA report, we hypothesize that (1a) students from the southern region do not use English as much as those in NCR and (1b) have a different attitude towards and perception of the language, (2a) that in-game performance and (2b) comprehension scores of south-based students are not as good as that of NCR students, but due to the prevalence of mobile games, students from both groups would have the same reception of the game in terms of their (3a) game-based learning engagement and (3b) intrinsic motivation.

This paper reports the testing of the mobile game *Learning Likha: Rangers to the Rescue* on students from southern Philippines and aims to answer the following questions:

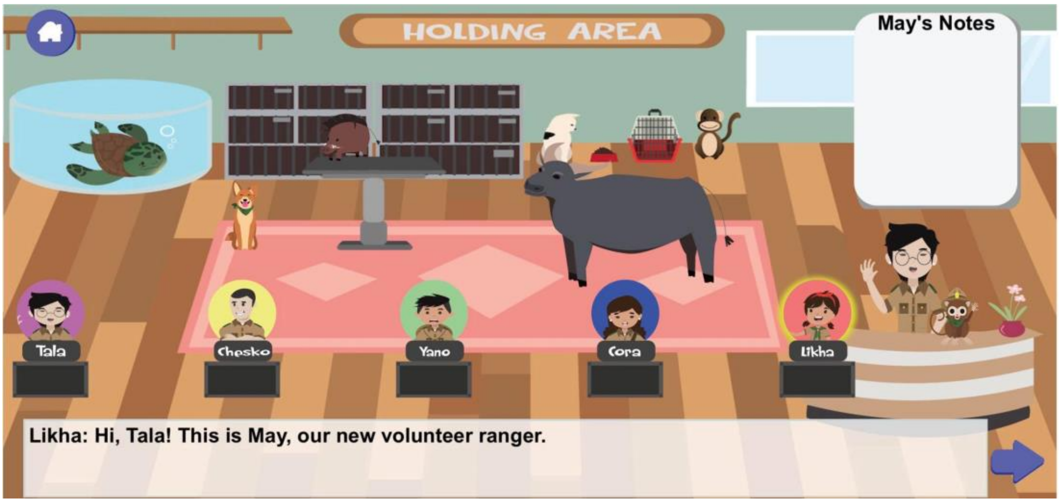
1) What are the differences in terms of usage, attitudes towards, and perceptions of the English language between students from southern Philippines and NCR?
2) How will the LLRR in-game performance, post-test comprehension scores, engagement, and motivation of students differ between the groups?

2. Narrative-Centered Digital Games and Learning Likha: Rangers to the Rescue

Narrative-Centered Digital Games (NCDGs) leverage interactive story scenarios through the use of characters and immersive plots combined with digital game environments to deliver educational content and problem-solving activities (Rowe, Shores, Mott, & Lester, 2010). The Narrative-Centered Learning Theory states that there are two ways in which a narrative can help motivate the learners: first, narrative text transports learners to another time and place that becomes real to them; and second, as learners interact with the material, they themselves become part of the narrative. What makes narrative an appealing approach is its capability to give meaningful structure where pedagogical objectives can be embedded into a coherent form that serves as a powerful motivating force for learners (Lester, Rowe, & Mott, 2013).

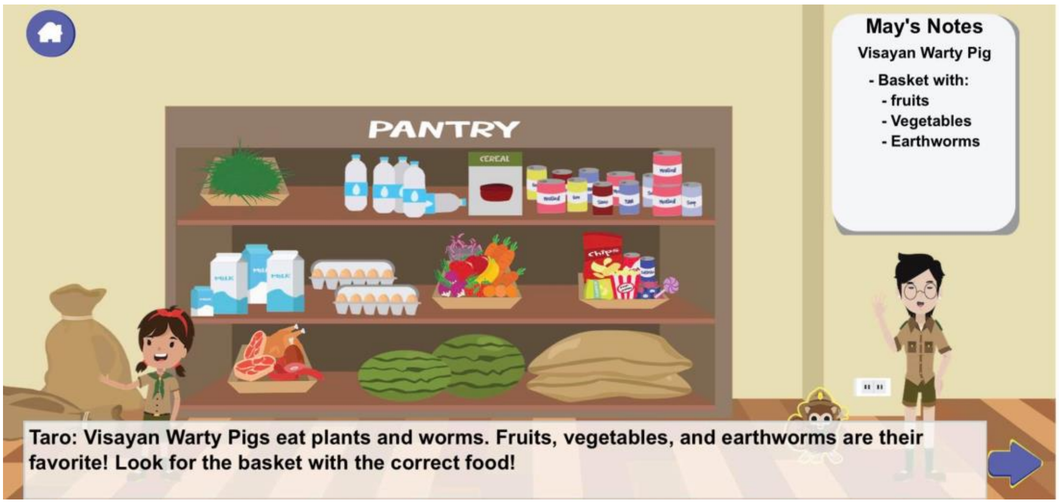
We attempt to use NCDGs for language learning (Williamson, 2009). To navigate through a narrative game, the player needs to listen to spoken instructions, read texts, and interpret visual cues. *Learning Likha: Rangers to the Rescue* (LLRR) (Agapito et al., 2020) is an NCDG that practices learners' ability to notice and understand details through written and oral language. Its secondary goals is to expose the learners to a variety of endangered species in the Philippines (Agapito et al., 2020). As the game starts, the player is introduced to the game's setting, gameplay, and objectives. The player is invited by Likha and her friend Taro the Tarsier to the Rescue Center in their town, Hiraya. The player is introduced to four other volunteer rangers, namely Tala, Chesko, Yano, and Cora, who need help to rescue endangered animals (see Figure 1).

After the player selects a ranger to help, spoken and written dialogues between Likha, Taro, and the ranger communicate what needs to be accomplished. These include a description of the endangered animal to be rescued or instructions of task(s) that the player needs to perform (see Figure 2). Feedback is given when the player makes an incorrect action and depending on the scenario, further instructions are provided for the player to complete the tasks successfully.



Likha: Hi, Tala! This is May, our new volunteer ranger.

Figure 1. Hiraya's Rescue Center's Holding Area where player selects one of the rangers.



May's Notes
Visayan Warty Pig
- Basket with:
  - fruits
  - Vegetables
  - Earthworms

Taro: Visayan Warty Pigs eat plants and worms. Fruits, vegetables, and earthworms are their favorite! Look for the basket with the correct food!

Figure 2. Sample dialogue that describes the one of the tasks.

The player needs to complete one or more tasks for a selected ranger to rescue a particular endangered animal. After which, the game gives a short explanation about the animal using both text and audio (see Figure 3). A full description of this game and its initial test results are available from Agapito et al., (2020).

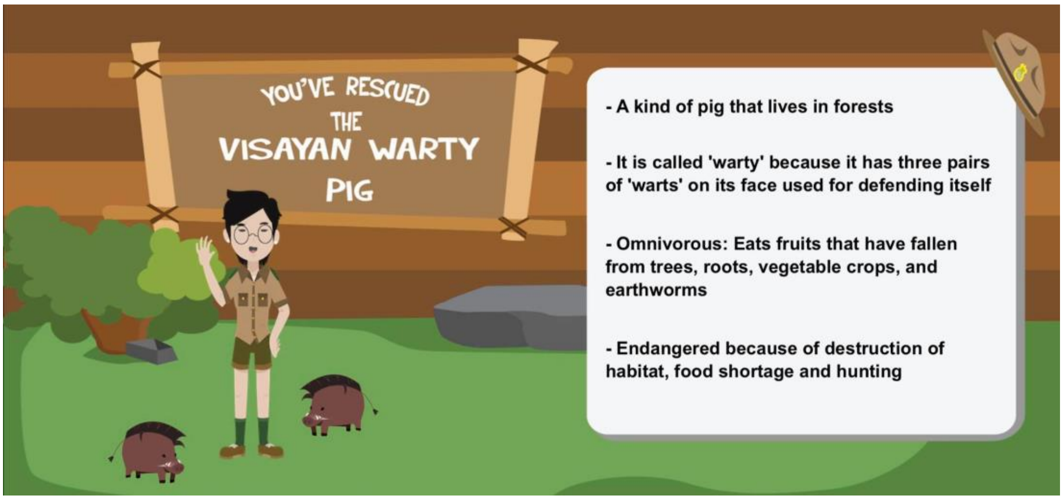

YOU'VE RESCUED THE VISAYAN WARTY PIG
- A kind of pig that lives in forests
- It is called 'warty' because it has three pairs of 'warts' on its face used for defending itself
- Omnivorous: Eats fruits that have fallen from trees, roots, vegetable crops, and earthworms
- Endangered because of destruction of habitat, food shortage and hunting

Figure 3. Description of the Visayan Warty Pig shown after the tasks have been completed.

3. Urban vs Rural English Language Learning in Other Countries

A study on Malaysian urban and rural students' attributions for success and failure in learning English as a second language (Gobel, Thang, Sidhu, Oon, & Chan, 2013) showed that urban and rural students had different attribution ratings for the success and failure for learning English as a second language. The urban group being more willing to attribute success to their own ability, effort, and study skills than students from the rural group. The rural group seems to attribute the failure more to the task they are given. Based on the data, the researchers hypothesized that the students in the urban group are more study-wise and confident as they have a greater belief in their own ability to take control of their successes in the language classroom.

In Indonesia, the work of Lamb (2012) offers strong evidence for the existence of regional differences in junior high school students' motivation to learn English, and in their progress. As an example, a young Indonesian's chance of achieving mastery of English depends mostly on where (s)he lives, as those who live in the central area of a city having a significant advantage over those from rural areas. They are also more likely to be motivated to learn English than their rural counterparts, in almost all aspects. Nevertheless, rural learners have shown to still have positive attitudes towards the language.

If these differences are not addressed, the gap between "the English-speaking 'have's" and the "non-English-speaking 'have-nots'" will continue to widen with the former living mainly in the urban areas, and the latter in the rural areas (Phillipson, 2012). If rural learners continue to get frustrated in their efforts to learn English, they may not be able to contribute to, or benefiting from, sectors of the local economy influenced by globalization.

4. Data Collection

Learning Likha: Rangers to the Rescue was tested on elementary students (grades 4, 5, and 6) from rural and urban public schools in Bukidnon, a province in southern Philippines. The mother tongue of these students is Cebuano. Since there were no face-to-face classes due to the pandemic, a field staff member was assigned to go to each of the participant's house to deliver the questionnaires and the mobile device for testing while observing the required safety protocols. The researcher then communicated with the participants over the phone for orientation and instructions.

The participants were given a demographics questionnaire to assess their level of access to mobile devices, their usage, attitude towards, and perceptions of the English language. They were given statements like "I speak English with my friends" and "I feel nervous when I need to speak English in class" and they indicated their level of agreement using a five-point Likert scale (1=Strongly Disagree to 5=Strongly Agree).

After playing LLRR, the participants were asked to answer a post-test on the different tasks and details of the endangered animals covered in the game to test their comprehension. They also answered the Game-Based Learning (GBL) Engagement Metric (Chew, 2017) adapted for LLRR to determine how engaged the students were with the game. They were given statements like "While playing *Learning Likha: Rangers to the Rescue*, I try my best to identify the details of the story" and they indicated their level of agreement using a five-point Likert scale (1=Strongly Disagree to 5=Strongly Agree). Next, they were given the Intrinsic Motivation Inventory (IMI) (Ryan, 1982) questionnaire with statements like "It was important for me to do well in *Learning Likha: Rangers to the Rescue*" and they indicated their level of agreement using a seven-point scale (1=Not at all true to 7=Very true).

The NCR data used in this study for the purposes of comparison is the same data from Agapito et al., (2020) which was earlier collected using the same protocol and questionnaires but in a face-to-face setting.

5. Results and Discussion

5.1 Profile of Participants and Attitude towards English

Out of the 90 participants from the south, only 36 (40%) have their own cellphone but 55 (61%) played mobile games by borrowing mobile phones from family and friends and 43 (48%) of them played educational games (see Table 1 for comparison with NCR).

Table 1. *Participants' Profile*

	Southern Region	NCR
Sex	Female = 43, Male= 47	Female = 14, Male = 13
Average age	10.79	10.30
Had their own mobile phone	36 (40%)	16 (59%)
Played mobile games	55 (61%)	23 (85%)
Played mobile educational games	43 (48%)	20 (74%)

The southern group had the smaller percentage of respondents who speak English at home (29%) and with their friends (34%) compared to NCR with 41% on both circumstances. When individual ratings were compared, no significant difference was found in terms of English usage between the groups which leads the researchers to reject hypothesis 1a. However, participants from NCR expressed more desire and enjoyment in learning and reading in English and they also have a higher perception on the importance of learning the language. More participants from the southern region find the language difficult to learn and that they feel nervous when they need to speak English in class (see Table 2). These findings support the hypothesis 1b that students from the south have a different attitude towards and perception of the language.

Table 2. *Attitude towards English: Southern Region and NCR Comparison*

Questions	South		NCR		t-value	p-value
	mean	sd	mean	sd		
1. I speak English at home.	2.74	1.14	3.04	1.26	1.130	0.260
2. I speak English with my friends.	2.88	1.16	3.07	1.30	0.741	0.460
3. I enjoy learning English.	3.76	1.08	4.26	0.93	2.177	0.031
4. I enjoy reading in English.	3.36	1.36	4.30	0.94	3.333	0.001
5. I find English difficult to learn.	3.26	1.26	2.59	1.06	-2.459	0.015
6. I feel nervous when I speak English in class.	3.48	1.25	2.70	1.24	-2.804	0.006
7. I want to learn to speak and read in English.	3.98	1.05	4.52	0.88	2.405	0.018
8. Learning English is important.	3.96	1.11	4.52	1.07	2.304	0.023

5.2 In-game Performance and Post-test Comprehension Scores

The in-game tasks in LLRR are performed by tapping on specific parts of the screen based on the given instructions. The player will not be able to proceed to the next scenario unless the task had been successfully performed. For each scenario, there is an ideal minimum tap count to perform the task (i.e. the player was able to perform the task with just the first instruction) and an ideal maximum tap count (i.e. the player had to reveal all the tips on how to perform the task). For example, in the Eagle 1 scenario, the minimum tap count is 4 composed of 1 tap on the eagle right after reading the displayed dialogue plus 3 taps required for the succeeding dialogues as the game transitions to the next scenario. The ideal maximum tap count is 7 composed of 3 taps on the dialogue arrow to reveal all hints for the task, then 1 tap on the eagle, then the 3 required arrow taps before the game transitions to the next scenario. To account for in-game performance, the players' tap count for each scenario was compared against the ideal maximum tap count. In this same context, the excess taps are considered as incorrect moves. It was also noted that no player has a tap count lower than the ideal maximum tap count for all scenarios.

To get a standardized in-game performance, we divided the ideal maximum tap count by the total number of the player's tap count (e.g. 74 ideal maximum tap count over the player's 135 taps results to a 54.81% in-game performance rating).

Comparing the in-game performance of the students from the southern region (*M* = 58.76%, *SD*=19.71%) to NCR students (*M* = 68.83%, *SD* = 15.99%) yields a significant difference, *t* (115) = -2.396, *p* = 0.009. This result supports hypothesis 2a as the in-game performance of students from the south are not as good as those from NCR.

The comprehension test given after the participants played the game was composed of 14 multiple-choice questions and 5 open ended questions for a total of 25 points. Only 42% of the students from the south while there were 85% from the NCR group who obtained a comprehension rating of 50% and above. A comparison of the post-test performance showed that NCR-based students (*M* = 72%, *SD* = 17%) had significantly higher comprehension ratings than those from the southern region (*M* = 45%, *SD* = 16%), *t* (115) = -7.535, *p* < .00001. This result is consistent with the hypothesis 2b as comprehension scores of south-based students are not as good as the NCR group.

5.3 Engagement and Intrinsic Motivation

The features for the analysis of engagement and motivation (see Table 3a and 3b) are adapted from the work of Moreno et al. (2019) on *Learning Likha: Music for the Fiesta*, the predecessor of LLRR.

Table 3a. *Description of engagement features from the GBL Engagement Metric (Moreno et al., 2019)*

Feature	Description
Behavior Engagement (BE)	Behavior engagement is a subcomponent of the GBL Engagement Metric. It refers to the actions a learner does which signals attentiveness to the game and engagement. This includes listening to instructions and problem solving.
Cognitive Engagement (CE)	Cognitive engagement is a subcomponent of the GBL Engagement Metric. It refers to the learners' experience of conceiving strategies and linking the activity to prior knowledge and skills.
Emotion Engagement EE)	Emotion engagement is a subcomponent of GBL Engagement Metric. It refers to the learner's physiological state, e.g. bored or having fun, while playing the game.

Table 4b. *Description of motivation features from Intrinsic Motivation Inventory (Moreno et al., 2019)*

Feature	Description
Enjoyment (En)	Enjoyment is a subcomponent of the IMI. It is the sustained interest of the learner while playing the game.
Effort (Ef)	Effort is a subcomponent of the IMI. It refers to the learner's self-reported estimate of how much effort and importance was placed in completing the game.

Perceived Competence (PC)	Perceived competence is a subcomponent of the IMI. It refers to the learner's perception of their own competence in completing in-game tasks.

The value for each feature was obtained from responses to the GBLE (scale of 1 to 5) and IMI (scale of 1 to 7) self-report questionnaires. The first analysis was a comparison of these features between the NRC and southern groups (see Table 5). There was no significant difference in cognitive engagement between the two groups as 93% of NCR-based participants and 91% from the southern group self-reported that they tried to use and apply what they have learned in class while playing the game, they considered asking questions when they didn't know what to do, and considered the game to have enough difficulty to challenge them. However, for all the other GBL and IMI features, there was a significant difference between the groups. The participants in the NCR group have significantly higher self-report ratings for the following: 1) behavior engagement, i.e. they tend to be more attentive, listened to instructions and tried their best to identify details of the story, 2) emotion engagement, i.e. they felt interested while playing, they look forward to finish the LLRR tasks, 3) enjoyment, i.e. they enjoyed and had fun and found the game interesting and not boring, 4) effort, i.e. they tried very hard to perform the tasks and it was important for them to do well, and 5) perceived competence, i.e. they thought they played LLRR pretty good and satisfied with their performance.

Table 5. *Engagement and Motivation: Southern Region and NCR Comparison*

Questions	South		NCR		t-value	p-value
	mean	sd	mean	sd		
Behavior Engagement (BE)	3.94	0.87	4.65	0.79	3.77	<0.001
Cognitive Engagement (CE)	4.08	0.74	3.96	0.72	-0.688	0.492
Emotion Engagement (EE)	3.99	0.83	4.65	0.79	3.627	<0.001
Enjoyment (En)	5.11	1.16	6.38	0.78	5.310	<0.001
Effort (Ef)	5.47	1.47	6.70	0.58	4.243	<0.001
Perceived Competence (PC)	3.43	1.48	6.27	0.66	9.602	<0.001

These differences in their GBL and IMI self-report ratings reject hypotheses 3a and 3b as their reception towards the game in terms of their (3a) game-based learning engagement and (3b) intrinsic motivation are different.

The second analysis was to investigate which features exhibited significant relationships with their post-test comprehension ratings, this was done for each group. The relationships between the GBL features, IMI features, and in-game ratings were checked using a series of Pearson's product-moment correlation coefficient (r). For the NCR data, results showed that only the GBL features (BE, CE, and EE) are highly correlated with each other, hence, the overall average engagement rating was instead used for a multiple linear regression. Using M5 Prime for feature selection, the regression result is shown in Table 5, effort had a slightly significant inverse relationship with comprehension while the overall engagement had a strong positive relationship with their English post-test comprehension rating.

Table 6. *NCR Group Comprehension Ratings: Multiple Linear Regression Coefficients*

	Coefficient	Std. Error	t-Stat	p-Value
Overall Engagement	0.118	0.039	3.035	0.006
Effort	-0.119	0.048	-2.500	0.020
Intercept	0.992	0.354	2.803	0.010

For the south group data, Pearson's product-moment correlation coefficient (r) showed that all engagement and motivation features were significantly correlated with each other. When the overall engagement average was tested with the overall motivation average, the relationship was still significant (r (88) = 0.409, p < 0.001). Thus, instead of doing a multiple linear regression for this group, we explored the relationship of comprehension ratings with their in-game performance, the remaining independent variable with no significant relationship with the other features. Result showed a significant correlation between comprehension and in-game performance, r (88) = 0.354, p<0.001.

These findings show that the relationships observed between engagement and motivation features, in-game performance, and the post-test comprehension ratings are different for each group.

6. Summary and Conclusion

It was expected that NCR-based participants use English more but results revealed that only less than half of the NCR group do so compared to about a third of the southern group and the difference was not significant. When it comes to their attitude towards the language, participants from NCR expressed more desire and enjoyment in learning and reading in English. They also have a higher perception on the importance of learning English. More participants from the southern group find the language difficult to learn and that they feel nervous when they need to speak English in class. These findings support the hypothesis that students from the south have a different attitude towards and perception of the language.

Learning Likha: Rangers to the Rescue received different responses from the two groups in terms of their in-game performance, game-based learning engagement and intrinsic motivation and these differences were reflected in the English comprehension post-test ratings.

The in-game performance of the participants in LLRR was measured by the maximum number of correct tap count over their total tap count. Result of the comparison showed that in-game performance of NCR-based participants are significantly higher than those from the south. This supports the hypothesis that the in-game performance of students from the south are not as good as those from NCR. The same gap was observed in the English comprehension post-test where NCR-based participants obtained significantly higher ratings which also supports the hypothesis that comprehension scores of southern-based students are not as good as the NCR group.

The differences are also consistent in terms of their engagement and motivation. For all the GBL and IMI features, except for cognitive engagement, the NCR group have significantly higher self-report ratings such that: they tend to be more attentive, listened to instructions and tried their best to identify details of the story; they felt interested while playing, they look forward to finish the LLRR tasks; they enjoyed and had fun and found the game interesting and not boring; they tried very hard to perform the tasks and it was important for them to do well; and they thought they played LLRR pretty well and were satisfied with their performance. These results lead the researchers to reject the hypothesis that reception of the game in terms of engagement and motivation are the same.

Finally, it was found that for the NCR group, comprehension rating had no significant relationship with their in-game performance but had a slightly significant inverse relationship with self-report effort exerted and a strong positive relationship with overall engagement. For the southern group, the participants' comprehension ratings were highly correlated with their in-game performance.

The differences between NCR and southern province students is representative of socio-economic differences between these groups. As English language proficiency can be a gateway skill that affects access to economic opportunities, improving English proficiency is one step towards improving socio-economic equity. Games such as LLRR are one of many materials that can potentially help educators reach these ends.

Acknowledgements

We thank the Ateneo de Manila University Research Council for the post-doctoral grant of MMPTF that made this study possible; the Ateneo Laboratory for the Learning Sciences of Ateneo de Manila University for the *Learning Likha: Rangers to the Rescue* game and for the mobiles phones used in the field test; and the students of the public schools in Bukidnon, Philippines who participated in this study.

References

Agapito, J. L., Manahan, D. M. A., Moreno, M. M. L., Beraquit, J. I., Herras, I. Y., Mora, K. A. C., … Rodrigo, M. M. T. (2020). Development and field testing of a narrative-centered digital game for english comprehension. *28th International Conference on Computers in Education, ICCE 2020*, 164–172. Asia-Pacific Society for Computers in Education.

EF EPI. (2020). *Education First English Proficiency Index*. Florida, USA. Retrieved from https://www.ef.com/wwen/epi/regions/asia/philippines/

Gobel, P., Thang, S. M., Sidhu, G. K., Oon, S. I., & Chan, Y. F. (2013). Attributions to success and failure in English language learning: A comparative study of urban and rural undergraduates in Malaysia. *Asian Social Science*, *9*(2), 53.

Lamb, M. (2012). A self system perspective on young adolescents' motivation to learn English in urban and rural settings. *Language Learning*, *62*(4), 997–1023.

Lema, K. (2017, November 9). Rise of the machines: Philippine outsourcing industry braces for AI. *Reuters*. Retrieved from https://www.reuters.com/article/us-philippines-economy-outsourcing-idUSKBN1D90BH

Lester, J. C., Rowe, J. P., & Mott, B. W. (2013). Narrative-centered learning environments: A story-centric approach to educational games. In *Emerging Technologies for the Classroom* (pp. 223–237). Springer.

Mariñas, J. (2021, March 2). 3 Reasons Why The Philippines is One of the Top English-Proficient Countries for Business. Retrieved April 28, 2021, from IT Outsourcing website: https://cloudemployee.co.uk/blog/it-outsourcing/why-philippines-for-business/

Moreno, M., Manahan, D., Fernandez, M., Banawan, M., Beraquit, J., Caparos, M., … Rodrigo, M. M. T. (2019). Development and Testing of a Mobile Game for English Proficiency Among Filipino Learners. *Proceedings of the 27th International Conference on Computers in Education*. Presented at the 27th International Conference on Computers in Education, Taiwan.

Ocumpaugh, J., Rodrigo, M. M., Porayska-Pomsta, K. K., Olatunji, U., & Luckin, R. (2018). Becoming Better Versed: Towards Design of Popular Music-based Rhyming Game for Disadvantaged Youths. *Proceedings of the 26th International Conference on Computers in Education*. Asia-Pacific Society for Computers in Education.

OECD. (2019). *Philippines—Country Note—PISA 2018 Results*.

Philippine Department of Education. (2019). *PISA 2018 National Report of the Philippines* [Department of Education Complex, Meralco Avenue, Pasig City Philippines]. Retrieved from https://www.deped.gov.ph/wp-content/uploads/2019/12/PISA-2018-Philippine-National-Report.pdf

Phillipson, R. (2012). Linguistic imperialism. *The Encyclopedia of Applied Linguistics*, 1–7.

Romero, P. (2018, February 22). Senate to probe declining English proficiency. *The Philippine Star*. Retrieved from https://www.philstar.com/other-sections/education-and-home/2018/02/22/1790069/senate-probe-declining-english-proficiency

Rowe, J. P., Shores, L. R., Mott, B. W., & Lester, J. C. (2010). Integrating learning and engagement in narrative-centered learning environments. *International Conference on Intelligent Tutoring Systems*, 166–177. Springer.

Shobikah, N. (2017). The Importance of English Language in Facing Asean Economic Community (AEC). *At-Turats*, *11*, 85–93.

Williamson, B. (2009). *Computer games, schools, and young people: A report for educators on using games for learning*. Futurelab Bristol.

Yiakoumetti, A., Evans, M., & Esch, E. (2005). Language awareness in a bidialectal setting: The oral performance and language attitudes of urban and rural students in Cyprus. *Language Awareness*, *14*(4), 254–260.

A Quasi-experimental Study of Chinese University English Learners' Engagement in a Flipped Classroom

Jingjing LIAN[ab*] & Jiyou JIA[b]
[a]*School of Humanities, Beijing University of Posts and Telecommunications, China*
[b]*Graduate School of Education, Peking University, China*
*lianjingjing@bupt.edu.cn

Abstract: The demand for improving the quality of undergraduate education in China has prompted the nation-wide implementation of blended teaching approach, including flipped classroom. However, there is a lack of studies examining the effect of flipped language classroom on students' behavioral, emotional, cognitive, and social engagement. To address this issue, this study examined the effect of a flipped English classroom on student engagement by adopting a pretest-posttest quasi-experimental design with mixed methods in a Chinese university. Data from 492 participants were collected and the results of the exploratory factor analysis (EFA) indicated that the instrument for measuring English learners' engagement was sufficiently reliable. The results of the analysis of covariance (ANCOVA) showed that students in the experimental group (n = 268) reported significantly higher behavioral, emotional, cognitive, and social engagement than students in the control group. Students' perceptions of their before-class and in-class learning experiences and their attitudes towards the flipped approach revealed some evidence for how and why flipped classroom could benefit language learners' engagement. The findings shed some light on future course design and the implementation of blended learning approach.

Keywords: Flipped classroom, engagement, English learners, quasi-experiment, ANCOVA

1. Introduction

Emerging technology applied to online learning has accelerated the adoption of blended learning in higher education. In China, the Ministry of Education issued the *Education Informatization 2.0 Action Plan* in 2018, aiming at developing a learner-centered education ecology integrated with information technology (Yan & Yang, 2020). Accordingly, Chinese universities are encouraged to build smart classrooms and adopt blended learning, such as flipped classroom approach to improve the quality of undergraduate education. Although the effect of flipped classroom has been widely discussed and examined since it gained its popularity in the United States (Bergmann & Sams, 2012), empirical research in the field of English language teaching remains limited (Turan & Akdag-Cimen, 2020), and there is a need for more rigorous research design to explore the effect of flipped classroom on students' learning outcomes (van Alten et al., 2019). To address this issue, this study conducted a quasi-experiment with mixed methods in an English course at a Chinese university to examine the effect of flipped classroom on university English learners' engagement.

2. Literature Review

2.1 Flipped Classroom

Flipped classroom is defined as a learning model in which lessons are delivered outside the classroom using instructional videos, leaving the in-class time for problem-solving and other activities (Bergmann & Sams, 2012). Since the work of Bergmann and Sams (2012) and other pioneers, flipped classroom has been gaining momentum and a vast number of studies have been conducted to investigate its pedagogical design, effectiveness, and influencing factors. A list of systematic reviews also emerged in recent years in an attempt to determine its advantages and challenges in various disciplines (e.g. Akçayır & Akçayır, 2018; O'Flaherty & Phillips, 2015; van Alten et al., 2019). Generally speaking, flipped classroom has been identified as an effective approach to improving teaching efficiency, students' satisfaction, engagement, and learning performance. Meanwhile, the successful implementation of flipped classroom also depends upon a wide range of factors concerning the out-of-class and in-class learning design (Lo, Hew, & Chen, 2017).

In the field of English language teaching, Turan and Akdag-Cimen (2020) conducted a systematic review and found that existing literature mostly showed positive effect of the flipped pedagogy. However, they pointed out that more experimental studies were needed and more qualitative findings were necessary to offer insights into the use of the flipped model in English classrooms. Since "flipped classroom" was introduced in China's *Guidelines on College English Teaching*, there has been a sharp increase of studies on flipped language classroom in Chinese universities (Qu, 2019). By conducting a content analysis of 42 Chinese research articles on flipped language classroom in higher education from 2014 to 2018, Qu (2019) suggested that there had been insufficient research on Chinese language learners and teaching assessment. Jiang et al. (2020) also maintained in their review that prevailing research was outcome-oriented without sufficient investigation into learners' perceptions and learning processes. Therefore, more rigorous experimental design and more qualitative inquiries are needed.

2.2 Engagement

Primarily a concept for understanding dropout and school completion (Finn, 1989), *student engagement* was defined by Christenson, Reschly, and Wylie (2012, vi) as "effortful learning through interaction with the teacher and the classroom learning opportunities". It has been widely acknowledged in the field of education that engagement is "the direct (and only) pathway to cumulative learning, long-term achievement, and eventual academic success" (Skinner & Pitzer, 2012, pp. 22-23), and has been recognized as a multifaceted construct. Based on extensive literature review, Fredricks, Blumenfeld, and Paris (2004) classified the existing measures of engagement into three dimensions: behavioral, emotional, and cognitive. According to Fredricks et al. (2016), behavioral engagement has been measured with items about attention, participation, concentration, and homework completion; emotional engagement is conceptualized as the presence of positive emotional reactions to teachers, peers, learning content, and classroom activities; cognitive engagement is defined in terms of using deep learning strategies, persistence, and self-regulated learning. In recent years, a fourth dimension – social engagement – was proposed to stress the importance of social interactions in learning, which was supported by Fredricks et al.'s (2016) qualitative study on math and science engagement. Meanwhile, Philp and Duchesne (2016) also theorized engagement as a multi-dimensional construct with the same four dimensions for second and foreign language learners. However, compared with the extensive research on engagement in educational psychology, there has been limited discussions of engagement among language educators. Instead, they traditionally focused more on *motivation* when examining language learners' commitment (Mercer & Dörnyei, 2020). Yet as the outward manifestation of motivation (Skinner & Pitzer, 2012), engagement offers a more practical approach to involving students in their language learning, especially in today's digital age where too many distractions might interfere with learners' effort even if they were motivated (Mercer & Dörnyei, 2020).

So far, many studies have explored whether flipped classroom could promote university students' behavioral, emotional, and cognitive engagement in various disciplines, such as in educational technology (Elmaadaway, 2018), computer science (Subramaniam & Muniandy, 2019), and language learning (Jamaludin & Osman, 2014), and yielded overall positive findings. However, these studies didn't adopt a pretest-posttest approach and only compared data after the flipped experiment. Moreover, some research on engagement in flipped English classrooms used the term "engagement" in a more general sense, representing students' active involvement with materials or activities without investigating it under the four-dimensional framework (e.g. Alsowat, 2016). Consequently, more pretest-posttest quasi-experimental research investigating the effect of flipped classroom on English language learners' behavioral, emotional, cognitive, and social engagement is needed, together with an attempt to understand students' perceptions of their flipped English learning.

3. Research Questions

This study aims at exploring an effective flipped approach to promoting English language learners' engagement by conducting a quasi-experiment in a university English course. Based on the literature review, two research questions were proposed:

(1) Can flipped classroom significantly improve university English learners' behavioral, emotional, cognitive, and social engagement?

(2) How did students in the flipped language classroom perceive their learning experiences?

4. Overview of the Research Design

This study was carried out in a freshman English course at a northern Chinese university in the fall semester of 2019. It was a compulsory course designed to improve non-English majors' general English skills. The course followed a textbook-based syllabus with units of various topics, such as the Internet, Ways of Learning, etc. In the course, students should learn articles of different styles and practice their listening, speaking, reading, and writing skills. Various technological tools such as multimedia system, smartphone applications, and online learning platforms were used to support a blended learning approach. The 16-week course lasted for 90 minutes each week with a class size of 60-70 students.

Adopting a pretest-posttest quasi-experimental research design, four classes were assigned to the experimental group, and another four classes to the control group. Due to management challenges, it was unfeasible to assign a single instructor to all the classes. Consequently, the two groups were taught by two female instructors with similar age, educational background, and teaching ability. Previous students' anonymous ratings of the two instructors were retrieved from the educational administration system and compared concerning their teaching skills, professional knowledge, class organization, and the way they inspired and encouraged students. No significant difference was identified.

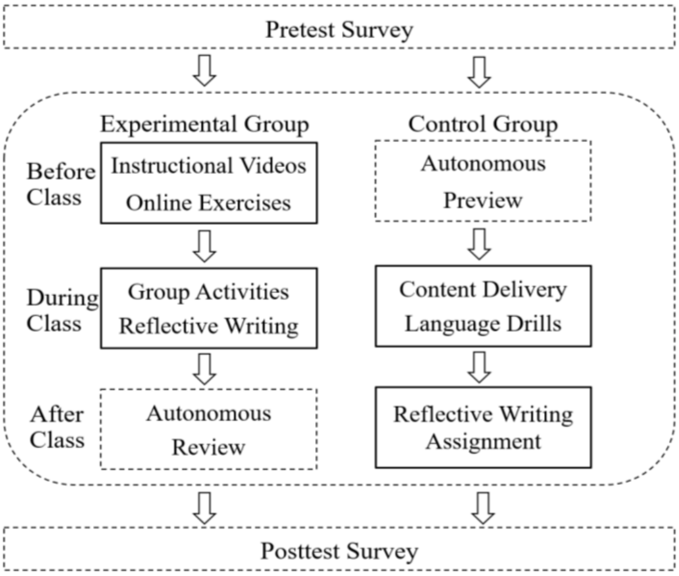


Figure 1. Procedure of the Teaching Experiment.

This study followed the design principles of flipped classroom proposed by Lo et al. (2017) and developed a flipped model for the English course. As shown in Figure 1, the experimental group adopted a flipped classroom approach, where the textbook-based instructions, such as analyzing the reading passages and explaining new words were moved to the *Blackboard* platform in the form of two 10-15 mins videos. The two instructors created the videos based on Mayer's (2014) cognitive theory of multimedia learning (CTML) and a previous survey on students' preferred elements in instructional videos. Students should watch the videos and complete the online exercises before attending the English class. During the class, the instructor first checked students' understanding of the pre-class knowledge with a small quiz and answered students' questions, then arranged learning activities such as reflective writing aimed at cultivating students' critical thinking. The control group followed the traditional blended teaching approach, where the instructor delivered the textbook-based knowledge and organized language drills and quizzes during the class, and left the reflective writing as after-class assignment. For both groups, digital courseware created by the textbook publisher was provided online to assist students' autonomous preview and review. Both instructors made it clear to students that class attendance and out-of-class assignments were mandatory, and their performance in class quizzes and activities accounted for 10% of the final grade. The two instructors also tried to observe the principle that the before-class learning load assigned to the experimental group should not be significantly different from the after-class learning assignment of the control group, in an effort to reduce the possible effect caused by extra workload in the experimental group (Turan & Akdag-Cimen, 2020). To examine the effect of flipped classroom on student engagement, two identical sets of survey were administered at the beginning and the end of the semester.

5. Method

5.1 Participants

The participants were 541 first-year undergraduate students from eight English classes, with four classes assigned to each group. Their English level was equivalent to CEFR B1-B2. Students of the two groups were from the same school (indicating same major and similar College Entrance Examination scores). The two groups of students attended the English course at the same time each week. All the students were engineering majors and they took the College English Placement Test when they entered the university. The scores of the test showed no significant difference between the two groups. Altogether 492 students completed both the pretest and posttest survey, with 268 students (77.6% males) in the experimental group and 224 students (69.2% males) in the control group. The average age of both groups is 18. According to the pretest survey, only 3.7% students (n = 10) in the experimental group and 4% students (n = 9) in the control group reported having prior learning experience in flipped classroom.

5.2 Instrument

The English Learning Engagement Scales (ELES) was used in this study to measure students' engagement in their English language learning. It was adapted from the Math and Science Engagement Scales developed by Fredricks et al. (2016), including four dimensions: Behavioral, Emotional, Cognitive, and Social Engagement. The items were modified by replacing "math and science" with "English". For example, when measuring students' emotional engagement, the original item "I look forward to math and science class." was modified to "I look forward to English class." All the items in the questionnaire were translated into Chinese by a professor of educational technology and an experienced English instructor and measured with a five-point scale, from 1 "I strongly disagree" to 5 "I strongly agree".

Besides the questionnaire, students from the experimental group were invited to post their flipped learning experiences on *Blackboard* at the end of the semester. The online instructions encouraged students to respond from three aspects: (1) their before-class learning experiences (including the quality of the online learning materials), (2) their in-class learning experiences, and (3) their perceptions of the flipped learning approach. This qualitative data was collected to triangulate the quantitative data by offering an in-depth understanding of students' learning processes and the effect of the flipped classroom on English learners' engagement.

5.3 Data Analysis

The quantitative data was analyzed using IBM SPSS 23.0. First, exploratory factor analysis (EFA) was performed to clarify the factors of the questionnaire, and the Cronbach's alpha was calculated to ensure the reliability of the instrument. Next, an analysis of covariance (ANCOVA) was employed to compare the engagement level of students in the two groups after the experiment. In terms of the qualitative data, content analysis (Popping, 2015) was adopted using NVivo 11 to analyze students' comments posted on *Blackboard*. Two researchers first read through all the data and noted that most students' responses consisted of three distinctive aspects as guided by the online instructions, i.e. their before-class learning experiences, their in-class learning experiences, and their attitudes towards the flipped learning approach. Consequently, the researchers first identified a comprehensive set of themes until no new theme emerged in the data and developed the coding scheme under the three major categories. Then they coded the responses independently to ensure investigator triangulation. The inter-coder agreement was greater than 80%, indicating that the procedures were sufficiently reliable. The two coders resolved disagreement by reviewing the responses. Themes irrelevant to the research question were not listed.

6. Results

6.1 Exploratory Factor Analysis of ELES

Exploratory factor analysis with principal components and varimax rotation was conducted using the pretest survey data (n = 492) to clarify the factors. Items with factor loadings lower than 0.40 or with multiple cross-loadings were excluded. Altogether 14 items were retained in ELES and grouped into Behavioral Engagement (BE, 4 items), Emotional Engagement (EE, 3 items), Cognitive Engagement (CE, 3 items), and Social Engagement (SE, 4 items). As shown in Table 1, the factor loadings of all the items ranged from 0.59 to 0.86. The total variance explained was 68.46%. The Cronbach's alpha for each factor ranged from 0.77 to 0.85, and the overall alpha was 0.91, indicating good internal consistency of the questionnaire. Among the four factors, BE displayed the highest mean score (4.20, SD = 0.62).

Table 1. *Descriptive Statistics, Factor Loadings, and Cronbach's Alpha Values for the ELES (N = 492)*

Factors	Items	Factor Loadings	M	SD
Behavioral	Mean = 4.20, SD = 0.62, Cronbach α = 0.82			
BE1	I stay focused in the English class.	0.74	4.05	0.84
BE2	I put effort into understanding learning content.	0.78	4.33	0.70
BE3	I keep trying even if something is hard.	0.67	4.26	0.74
BE4	I complete my learning tasks on time.	0.74	4.18	0.81
Emotional	Mean = 3.51, SD = 0.86, Cronbach α = 0.85			
EE1	I enjoy attending English class.	0.86	3.42	1.03
EE2	I look forward to English class.	0.86	3.24	1.03
EE3	I go to the class with good expectations when I know what to learn.	0.64	3.86	0.89
Cognitive	Mean = 3.48, SD = 0.80, Cronbach α = 0.77			
CE1	I think about different ways to complete the English assignment.	0.68	2.96	1.06
CE2	I try to connect what I am learning to things I have learned before.	0.78	3.75	0.93
CE3	I try to apply what I have learned in the English class.	0.71	3.71	0.92
Social	Mean = 3.88, SD = 0.72, Cronbach α = 0.84			
SE1	I try to understand other people's ideas in English class.	0.59	3.89	0.81
SE2	I try to learn with others who can help me in English.	0.80	3.92	0.89
SE3	I cooperate with others in English class.	0.83	3.89	0.85
SE4	I share my ideas when working with others in English class.	0.79	3.79	0.92

6.2 ANCOVA Results of Student Engagement

To examine whether there was a significant difference in student engagement between the two groups after the experiment, an analysis of covariance (ANCOVA) was performed using the posttest scores of ELES as dependent variables, the pretest scores of ELES as covariates, and the teaching approach as an independent variable. The four factors of engagement in ELES were analyzed separately. The homogeneity of slopes assumption in ANCOVA was tested, and the results showed no significant interaction between the teaching approach and each of the engagement factors (“BE” $F = 0.061$, $p > 0.05$; “CE” $F = 0.008$, $p > 0.05$; “EE” $F = 3.798$, $p > 0.05$; “SE” $F = 0.115$, $p > 0.05$), indicating that the assumption was met. The detailed descriptive statistics and ANCOVA summary are shown in Table 2.

The results of ANCOVA indicated that students in the experimental group reported significantly higher scores in their engagement than students in the control group, including their BE ($F = 28.58$, $p < 0.001$), EE ($F = 24.79$, $p < 0.001$), CE ($F = 15.52$, $p < 0.001$), and SE ($F = 36.83$, $p < 0.001$). As shown in Table 2, the average pretest scores of BE in both groups were the highest among the four factors, suggesting that when students just entered university, they reported quite high behavioral engagement in their previous English classes. However, after one-semester English course in university, the adjusted mean of the control group’s BE reduced to 3.67, and there was a slight decrease (adjusted M = 3.95) in the experimental group’s BE. However, in terms of Emotional, Cognitive, and Social Engagement, the adjusted mean scores of the experimental group were all higher than its pretest scores, whereas there was a general decrease in the adjusted mean scores of the control group.

Table 2. *Descriptive Statistics of Students’ Pretest and Posttest Scores and ANCOVA Summary of Students’ Engagement*

	Pretest		Posttest		Univariate ANCOVA			
	M	SD	M	SD	M (adjusted)	SE	*F value*	η^2
Behavioral Engagement (BE)								
Experimental Group	4.12	0.65	3.92	0.61	3.95	0.036	28.58^{***}	0.055
Control Group	4.31	0.58	3.71	0.66	3.67	0.039		
Emotional Engagement (EE)								
Experimental Group	3.39	0.85	3.78	0.73	3.82	0.041	24.79^{***}	0.048
Control Group	3.64	0.86	3.57	0.77	3.52	0.045		
Cognitive Engagement (CE)								
Experimental Group	3.38	0.80	3.57	0.73	3.61	0.039	15.52^{***}	0.031
Control Group	3.59	0.79	3.43	0.72	3.38	0.043		
Social Engagement (SE)								
Experimental Group	3.79	0.69	3.83	0.69	3.87	0.040	36.83^{***}	0.070
Control Group	3.98	0.73	3.55	0.72	3.51	0.043		

Note. Experimental Group n = 268, Control Group n = 224. ***p<0.001

6.3 Students’ Perceptions of the Flipped Classroom

Altogether 139 students in the experimental group posted their comments. Themes concerning students’ before-class and in-class learning experiences, and their attitudes towards the flipped classroom were coded. The frequency and the translated answer example of the most representative themes relevant to the research question are presented in Table 3, which offered a qualitative understanding of the ANCOVA results.

To begin with, students in the experimental group generally spoke highly of the instructional videos. They commented that the videos were of high quality, offering thorough explanations of textbook knowledge, including new vocabulary and article understanding. They could easily access and review those videos. In addition, the before-class online exercises helped deepen their understanding of the learning content, allowing them to get better prepared before attending English classes. They considered the exercises to be easy to deal with, and some students suggested adding more exam-oriented exercises. In terms of in-class learning, students showed overall positive attitudes towards group activities, through which they engaged in topic-related discussions, writing tasks, and group presentations. They noted that those activities had improved their collaborative and communicative skills. Most students reported that they enjoyed the learning process and could stay focused throughout the class. However, a few students pointed out that they were not accustomed to group activities. "*Why do I have to learn with others?*" One student commented. Some asked for more teacher's instructions on testing skills and English learning methods in class.

Table 3. *Students' Perceptions of the Flipped English Classroom (N =139)*

Theme	Frequency	Example
Before-class learning		
High quality instructional videos	101	*"The instructional videos are of high quality and interesting, with clear explanations of the textbook content."*
Helpful online exercises	54	*"The exercises helped deepen my understanding of the words and sentences in the textbook."*
Suggestions for improvement	32	*"Maybe the teacher could add more exercises relevant to the College English Test."*
In-class learning		
Rich and colorful	52	*"The group activities were rich and exciting. They not only enhanced the awareness of group cooperation, but also made the class more engaging."*
Efficient arrangement	34	*"The class was well-arranged. We first reviewed the textbook knowledge, then participated in topic-related activities. I could always stay attentive in the class."*
More instructions needed	21	*"I hope the class could provide more instructions on skills in English tests."*
Not all students were active	13	*"Some students were inactive in class, maybe due to their poor English skills or lack of interest in English learning."*
Attitudes towards the flipped learning		
Supportiveness and acceptance	63	*"I totally support this innovative approach. We have very few English classes now. If we just listen to the teacher lecturing throughout the class, the efficiency is very low."*
Improved autonomous learning	44	*"It (flipped classroom) allowed me to learn at my own pace and enhanced my autonomous learning."*
Improved motivation	32	*"The class provided more opportunities for us to express ourselves in English. I feel more motivated to learn English well."*
Changes in learning conceptions	26	*"This teaching approach has changed my understanding of English learning. Instead of passing exams, we are now learning English for communicative purposes."*
Challenges in adaptation	17	*"It took me almost a month to get used to this teaching approach."*

In general, most students acknowledged the value of the flipped approach. Although many of them mentioned that it was the first time for them to experience flipped classroom, they considered this approach as more effective in knowledge delivery and fostering students' active learning. The structured guidance on *Blackboard* and teacher's online support allowed students to learn at their own pace, which cultivated their autonomous learning. Some students mentioned that the flipped approach had changed their English learning conceptions – learning English was no longer simply for passing exams and getting a high score. Now they were more motivated to use English for communicative purposes. Despite the overall positive feedback, there were a few students struggling to adapt to the flipped approach. They said that they couldn't manage their time well and were always "*chasing deadlines*". Besides, many students expressed concerns about how to learn English well, as there were fewer English classes each week in university than in high school, whereas their major-related courses such as linear algebra and physics consumed much of their time and energy.

7. Discussion

This study examined the effect of flipped classroom on university English learners' engagement by carrying out a quasi-experiment in a Chinese university. The results of the EFA of ELES validated the survey instrument of English learners' engagement, and the ANCOVA analysis showed that after the experiment, students in the experimental group on average reported significantly higher behavioral, emotional, cognitive, and social engagement than students in the control group. The qualitative findings of students' overall positive attitude revealed some evidence that could explain and support the higher engagement level in the experimental group.

7.1 The Effect of Flipped Classroom on Behavioral Engagement

Students in the experimental group reported significantly higher behavioral engagement than students in the control group after the one-semester English course. This was consistent with previous findings that the flipped approach increased attendance and retention rate (Karabulut-Ilgu, Cherrez, & Jahren, 2018). It also proved that the mandatory requirement of pre-class learning and class attendance in the flipped classroom is necessary (He, Holton, Farkas, & Warschauer, 2016). However, the mean scores of students' posttest behavioral engagement in both groups were lower than those of their pretest. One possible explanation is that the participants might no longer treat English learning as the top priority after they entered into university, where they had only two hours of English classes once a week, and more major-related courses took up much of their effort. Under the circumstances, the average behavioral engagement of students in the control group reduced considerably, whereas the mean score of students' behavioral engagement in the experimental group didn't show significant change, indicating the effect of flipped classroom on maintaining students' attention and involvement in English learning. Meanwhile, it should be noted that the successful implementation of the flipped classroom relies on technological affordances and high-quality out-of-class materials (Akçayır & Akçayır, 2018). To engage learners behaviorally, the technological tools should be easy to use and accessible to all learners. And the instructor should manage students' transition to the flipped approach by demonstrating the learning process with technology and offering instant guidance and support.

7.2 The Effect of Flipped Classroom on Emotional Engagement

Students in the experimental group generally exhibited higher emotional engagement than students in the control group after the experiment, and their average posttest emotional engagement level was higher than that of their pretest. Most of them found the flipped class rich, colorful, and exciting, and they enjoyed attending the classes. Their comments revealed that although a few students took some time to adapt to this approach, most of them held supportive attitudes at the end of the semester and acknowledged its value in improving class efficiency and providing opportunities for communication. In many Chinese universities, the class size of the English course for non-English-major undergraduates was still quite large (60-70 students), and teacher-centered instructions took up much of the class time (Chen & Yu, 2019). Consequently, it was difficult for students in the control group to relate themselves to the teacher, the learning content, and their peers, which might lead to reduced emotional engagement. In comparison, students in the flipped classroom were more motivated with more opportunities to get involved in class activities. The result of the increased emotional engagement was in line with previous findings of increased student satisfaction with the flipped approach (O'Flaherty & Phillips, 2015).

7.3 The Effect of Flipped Classroom on Cognitive Engagement

Compared with students in the control group, students in the flipped classroom showed significantly higher cognitive engagement. Students generally spoke highly of the instructional videos and found them easy to understand and fun to watch, which might have better prepared them for higher-order learning activities in the class. Some of them commented that the course was well-organized, and they could stay attentive in class and apply what they had learned to group discussions and writing tasks, indicating deep learning strategies. From this perspective, clear guidance and explicit requirement from the teacher might have had some effect on students' self-regulation (Shyr & Chen, 2018). Previous research has identified that failures in flipped classroom could be attributed to poor-quality learning materials or poorly-arranged classroom activities (Akçayır & Akçayır, 2018). Consequently, educators should carefully design the flipped course to cognitively engage students. To begin with, the quality of the instructional videos played a vital role for effective flipped learning. The video producer could conduct a needs analysis among students and observe Mayer's (2014) CTML in order to reduce learners' cognitive load. Next, exercises directly linked to the video content should be offered online to consolidate students' learning. Besides, a short and easy quiz checking students' understanding of the pre-class knowledge is encouraged at the beginning of the class. After that, learning activities that emphasize communicative and higher-order thinking skills could be carried out based on the pre-class knowledge. To sum up, to genuinely "flip" the classroom, the pre-class and in-class activities should be carefully integrated with clear guidance to foster students' cognitive engagement. The positive findings indicated that Lo et al.'s (2017) design principles for mathematics flipped classrooms that this study had followed might also apply to language classrooms.

7.4 The Effect of Flipped Classroom on Social Engagement

Students in the flipped classroom reported significantly higher social engagement than students in the control group. They had more opportunities to work with peers, share and understand each other's ideas through group activities. However, as freshmen who just got into university after the College Entrance Examination, some students still showed inclination towards exam-oriented teaching approach and demanded more teacher instructions in class. This finding echoed Webb, Doman, and Pusey's (2014) survey that students in the flipped group still clung to in-class teacher instructions. Therefore, a few of them were not accustomed to suddenly becoming the center of the classroom and remained passive in the face of classroom activities. Nevertheless, the qualitative results showed that the flipped classroom allowed students to gradually change their conceptions of English learning and accept the importance of enhancing their English communicative skills through collaborating with peers.

8. Conclusion

This study conducted a pretest-posttest quasi-experiment in a Chinese university English course and examined the effect of flipped classroom on student engagement. The results showed that the flipped language classroom has significantly promoted English learners' engagement. And students' comments on their learning experiences revealed some evidence for the positive effect of the flipped classroom. In particular, the results identified the crucial role of the quality of out-of-class learning materials and the integration of pre-class and in-class learning tasks in engaging students behaviorally, emotionally, cognitively, and socially. This research has contributed to the understanding of effective approaches to promoting students' engagement. The design principles followed by this study could apply to flipped classrooms of other disciplines, and also shed some light on effective organization of online learning or other blended learning model.

This study has several limitations. To begin with, all the participants were engineering majors in a Chinese university, so the findings may not be fully generalizable. Besides, the English instructors in the two groups were not the same. Even though no significant difference was identified in their teaching style and professional skills and they kept their teaching content in sync, some of their individual characteristics might still lead to differences in student engagement. Future quasi-experimental study should ensure the same instructor for both groups. Furthermore, students' English proficiency after the experiment was not compared due to the lack of a validated large-scale English test. Studies examining the effect of flipped classroom on students' higher-order thinking skills and communicative language skills are still greatly needed.

Acknowledgements

We would like to thank all the people who helped review this paper.

This paper is supported by Beijing University of Posts and Telecommunications 2021 Teaching Reform Project "High-Quality Courses".

References

Akçayır, G., & Akçayır, M. (2018). The flipped classroom: A review of its advantages and challenges. *Computers & Education*, *126*, 334-345.

Alsowat, H. (2016). An EFL flipped classroom teaching model: Effects on English language higher-order thinking skills, student engagement and satisfaction. *Journal of Education and Practice, 7*(9), 108-121.

Bergmann, J., & Sams, A. (2012). *Flip your classroom: Reach every student in every class every day*. Eugene, Or: International Society for Technology in Education.

Chen, W., & Yu, S. (2019). Implementing collaborative writing in teacher-centered classroom contexts: student beliefs and perceptions. *Language Awareness, 28*(4), 247-267.

Christenson, S. L., Reschly, A. L., & Wylie, C. (Eds.). (2012). *Handbook of research on student engagement*. Springer Science & Business Media.

Elmaadaway, M. A. N. (2018). The effects of a flipped classroom approach on class engagement and skill performance in a blackboard course. *British Journal of Educational Technology, 49*(3), 479-491.

Finn, J. D. (1989). Withdrawing from school. *Review of Educational Research, 59*, 117-142.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: potential of the concept, state of the evidence. *Review of Educational Research, 74*, 59-109.

Fredricks, J. A., Wang, M. T., Schall Linn, J., Hofkens, T. L., Sung, H. C., Parr, A. K., & Allerton, J. J. (2016). Using qualitative methods to develop a survey measure of math and science engagement. *Learning and Instruction, 43*, 5-15.

He, W., Holton, A., Farkas, G. A., & Warschauer, M. (2016). The effects of flipped instruction on out-of-class study time, exam performance, and student perceptions. *Learning and Instruction, 45*, 61-71.

Jamaludin, R., & Osman, S. Z. M. (2014). The Use of a Flipped Classroom to Enhance Engagement and Promote Active Learning. *Journal of Education and Practice, 5*(2), 124–131.

Jiang, M. Y. C., Jong, M. S. Y., Lau, W. W. F., Chai, C. S., Liu, K. S. X., & Park, M. (2020). A scoping review on flipped classroom approach in language education: challenges, implications and an interaction model. *Computer Assisted Language Learning*, 1-32.

Karabulut-Ilgu, A., Cherrez, N. J., & Jahren, C. T. (2018). A systematic review of research on the flipped learning method in engineering education. *British Journal of Educational Technology*, *49*(3), 398-411.

Lo, C. K., Hew, K. F., & Chen, G. (2017). Toward a set of design principles for mathematics flipped classrooms: a synthesis of research in mathematics education. *Educational Research Review, 22*, 50-73.

Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.). *The Cambridge handbook of multimedia learning*: Second edition (pp. 43-71). New York: Cambridge University Press.

Mercer, S., & Dörnyei, Z. (2020). *Engaging language learners in contemporary classrooms*. Cambridge University Press.

O'Flaherty, J., & Phillips, C. (2015). The use of flipped classrooms in higher education: A scoping review. *The Internet and Higher Education, 25*, 85–95.

Philp, J., & Duchesne, S. (2016). Exploring engagement in tasks in the language classroom. *Annual Review of Applied Linguistics, 36*, 50-72.

Popping, R. (2015). Analyzing open-ended questions by means of text analysis procedures. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 128*(1), 23-39.

Qu, S. (2019). A Content Analysis of Researches on EFL Flipped Classroom in China's Universities. *Technology Enhanced Foreign Language Education* [Chinese Journal], *187*(3), 62-69.

Shyr, W. J., & Chen, C. H. (2018). Designing a technology-enhanced flipped learning system to facilitate students' self-regulation and performance. *Journal of Computer Assisted Learning, 34*(1), 53-62.

Skinner, E. A., & Pitzer, J. R. (2012). Developmental dynamics of student engagement, coping, and everyday resilience. In *Handbook of Research on Student Engagement* (pp. 21-44). Springer, Boston, MA.

Subramaniam, S. R., & Muniandy, B. (2019). The effect of flipped classroom on students' engagement. *Technology Knowledge & Learning, 24*, 355-372.

Turan, Z., & Akdag-Cimen, B. (2020). Flipped classroom in English language teaching: a systematic review. *Computer Assisted Language Learning, 33*(5-6), 590-606.

van Alten, D. C., Phielix, C., Janssen, J., & Kester, L. (2019). Effects of flipping the classroom on learning outcomes and satisfaction: A meta-analysis. *Educational Research Review, 28*, 100281.

Webb, M., Doman, E., & Pusey, K. (2014). Flipping a Chinese university EFL course: What students and teachers think of the model. *The Journal of Asia TEFL, 11*(4), 53-87.

Yan, S., & Yang, Y. (2020). Education Informatization 2.0 in China: Motivation, Framework, and Vision. *ECNU Review of Education*, 1-19.

Integrating E-learning into Self-regulated Learning Instruction: A Holistic Flipped Classroom Design of a Classical Chinese Reading Intervention Program

Kit Ling LAU
The Chinese University of Hong Kong, Hong Kong
*dinkylau@cuhk.edu.hk

Abstract: The paper introduces the design of an intervention program which integrates in-class self-regulated learning (SRL) instruction and out-of-class e-learning activities based on the theoretical model of flipped classroom to enhance Hong Kong junior secondary school students' learning of Classical Chinese (CC) reading. Four instructional principles of SRL instruction, namely task nature, teacher support, student autonomy, and mastery-oriented evaluation, are used to guide the design of learning activities. To compensate for the common problem of lacking teacher support in e-learning, in-class instruction will help students establish a good foundation for CC reading and extend their knowledge and skills learned in out-of-class e-learning activities. To change students' reliance on their teachers when learning CC reading, out-of-class e-learning activities will provide opportunities for students to practice their self-regulatory skills independently. While previous studies have demonstrated the benefits of SRL instruction and using technology to enhance language learning, this intervention program makes a first attempt to integrate SRL instruction and e-learning into CC reading instruction.

Keywords: flipped classroom, self-regulated learning, technology enhanced language learning

1. Introduction

Both self-regulated learning (SRL) and e-learning are effective forms of learning (Beetham, & Sharpe, 2019; Schunk & Greene, 2018). Previous experiences of implementing SRL instruction in Hong Kong indicate that teachers and students are less favorable to instructional principles that place more emphasis on student autonomy (Lau, 2011; 2013). Classical Chinese (CC) is a written form of old Chinese. Due to the difficulties students experience in CC reading, it is more demanding for them to develop SRL in CC reading. Against this background, an intervention program which integrates in-class SRL instruction and out-of-class e-learning activities based on the theoretical model of flipped classroom is proposed to enhance Hong Kong students' learning of CC reading in a holistic approach.

2. Background and the Theoretical Foundation of the Intervention Program

2.1 Using SRL Instruction to Enhance Classical Chinese Learning

SRL represents an effective form of learning that involves metacognition, motivation, and strategic actions (Zimmerman, 2000). Zimmerman (Schunk & Zimmerman, 1997; Zimmerman, 2000) has suggested that learners develop self-regulatory skills as they move through four levels: observational, emulation, self-controlled, and self-regulation level. Social cognitive theorists emphasize the importance of environmental factors in shaping students' approaches to learning. Based on the findings of previous studies (e.g., Housand & Reis, 2008; Lombaerts, Engels, & van Braak, 2009; Perry, 1998; Stoeger, Sontag, & Ziegler, 2014; van Grinsven & Tillema, 2006), the author developed a "TSAE" framework to group different effective instructional practices into four principles. "T" refers to the nature of the instructional "Task," which includes using direct strategy instruction and open tasks to enhance students' reading ability, facilitate strategy use, and motivation. "S" represents the teacher's "Support," which scaffolds students' development through various levels of self-regulatory skills. "A" characterizes the degree of "Autonomy" that students have in controlling their learning. Finally, "E" refers to the adoption of mastery-oriented "Evaluation" practices and student-involved assessment. In reference to the four developmental stages of SRL, the principles of "T" and "S" play a more prominent role in supporting students toward establishing a solid foundation of learning in the first two levels of SRL. When they attain the two higher levels of SRL, the principles of "A" and "E" are emphasized to offer a high degree of autonomy for students to learn independently.

The TSAE framework was applied in supporting teachers in implementing SRL instruction in Chinese language classes (Lau, 2011; 2013). While these studies supported the positive impacts of SRL instruction on student learning, teachers found it easier to apply the principles with more emphasis on the role of the teacher ("T" and "S") than giving a high degree of control to students ("A" and "E"). Students were also found to be easily failed to read efficiently without teachers' constant support. Recently, the author designed a CC reading intervention program based on the TSAE framework (Lau, 2020). Although learning to read CC texts is a core component of the Chinese language curriculum in Chinese societies, Chinese students' CC reading performance and motivation is unsatisfactory due to the linguistic differences between CC and modern Chinese and their insufficient background knowledge of ancient Chinese culture (Chi & Chiou, 2015; Lau, 2017; 2018). The effectiveness of the SRL-based CC reading intervention program was evaluated in a quasi-experimental study. The study's findings indicated that the intervention program was more effective than the traditional teacher-centered instruction in enhancing students' prior CC knowledge and comprehension. However, no significant changes in their strategy use and motivation were found in the quantitative data. While these findings support the feasibility of using SRL instruction to enhance students' CC reading, the intervention effects might be limited by students' weak foundation and its short duration.

2.2 Using E-learning to Facilitate SRL and Classical Chinese Learning

With the rapid development of technology, e-learning has been proposed to be an effective way to promote SRL (Anderton, 2006; Johnson & Davies, 2014; Narciss et al., 2007). Since e-learning provides autonomy for students to control their learning (Lee & Tsai, 2011; Sletten, 2017), adding e-learning activities in SRL instruction should be a good direction to increase the degree of "A" and "E" principles and change Chinese students' reliance on their teachers. Previous studies on TELL have also supported the effectiveness of using e-learning in helping students cultivate positive attitudes toward language learning and improving their performance in learning English (Hsieh et al., 2017; Turan & Akdag-Cimen, 2019) and Chinese as a second language (Tseng et al., 2018; Wang et al., 2018). Although previous e-learning studies were seldom conducted on CC learning, the promising results of TELL studies suggest that e-learning have considerable potential to facilitate CC learning. For example, teaching videos can be used to support students' learning of CC linguistic knowledge and vocabulary (Chen et al., 2017). Digital games can be designed for students to practice various CC reading skills (Hwang et al., 2012; Lin et al., 2020). The interactive and playful features of e-learning materials and activities can improve students' motivation in learning CC reading (De Grove et al., 2012).

Despite its many advantages, e-learning can be a double-edged sword for learning success (Lee & Tsai, 2011). While e-learning provides a high degree of autonomy for students to practice SRL, it is more demanding than the traditional learning environment. Students must possess adequate self-regulatory skills to learn effectively in an online environment (Anderton, 2006; Lee & Tsai, 2011; Sletten, 2017). To maximize the benefits and compensate for the limitations of SRL and e-learning, flipped classroom (FC) is adopted as the theoretical framework of the CC reading intervention program to combine these two types of effective learning to achieve a holistic instructional design. FC is a form of blended learning which reverses the traditional teacher-centered classroom into student-centered learning by having students study content material prior to class through online learning to free up in-class time for more interactive and higher-level learning activities (Berrett, 2012; Fulton, 2012). Compared with solely online learning, FC places more emphasis on the teacher's role as a facilitator to enable students to perform SRL (Johnson & Davies, 2014; Sletten, 2017). While out-of-class e-learning activities of FC have always been criticized as focusing on low-level learning tasks (Blau &

Shamir-Inbal, 2017; Lo & Hew, 2017), this problem can be avoided by applying the SRL instructional principles in designing the out-of-class e-learning activities.

3. Instructional Design of the Intervention Program

3.1 Theoretical Framework of the Program Design

By adding a new component of out-of-class e-learning activities, this paper proposes a new CC reading intervention program aiming to extend the author's previous intervention study (Lau, 2020) on using SRL instruction to enhance Hong Kong students' CC learning. The design of this new intervention program refers to the four-level model of SRL development, the TSAE framework of SRL instruction, studies on TELL and CC reading reviewed above. A "Re-designed model of flipped learning (RDFC model)" proposed by Blau & Shamir-Inbal (2017) is used as a holistic framework to integrate different types of in-class instruction and out-of-class e-learning activities in the program. Differing from the traditional FC model in which the learning of new content mostly occurs through pre-class video watching, this model promotes students' active learning in both in-class and out-of-class settings, with knowledge construction taking place before, during, and after the lessons (Figure 1).

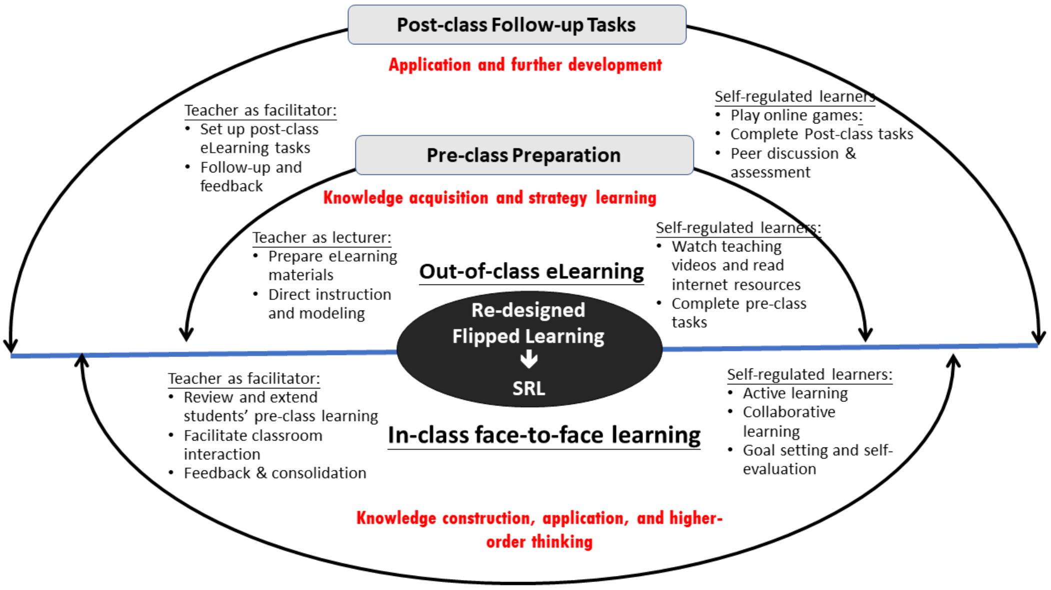

Post-class Follow-up Tasks
Application and further development
Teacher as facilitator:
• Set up post-class eLearning tasks
• Follow-up and feedback
Self-regulated learners:
• Play online games:
• Complete Post-class tasks
• Peer discussion & assessment
Pre-class Preparation
Knowledge acquisition and strategy learning
Teacher as lecturer:
• Prepare eLearning materials
• Direct instruction and modeling
Self-regulated learners:
• Watch teaching videos and read internet resources
• Complete pre-class tasks
Out-of-class eLearning
Re-designed Flipped Learning
SRL
In-class face-to-face learning
Teacher as facilitator:
• Review and extend students' pre-class learning
• Facilitate classroom interaction
• Feedback & consolidation
Self-regulated learners:
• Active learning
• Collaborative learning
• Goal setting and self-evaluation
Knowledge construction, application, and higher-order thinking

Figure 1. The Theoretical Framework of the Intervention Program (adapted from the RDFC model).

3.2 Description of the Instructional Design

The intervention program will be integrated into the experimental schools' regular Chinese language lessons for two years. With reference to the four developmental stages of SRL, the program is divided into four phases. In each phase, students learn new knowledge and strategies through pre-lesson e-learning activities to free up in-class instruction time for more interactive and higher-level learning activities. Post-class e-learning activities are designed to facilitate students' application and further development of their learned knowledge and strategies through more advanced learning tasks. The importance of different SRL instructional principles is adjusted according to students' developmental stages of SRL skills.

3.2.1 Phase 1: Observation Level

SRL Instructional Principles: Focus on building up students' foundation on CC reading and emphasis the supportive role of the teacher.

- **T**: All learning materials are selected based on a humanistic theme of filial piety to enhance students' background knowledge of CC reading. Different types of CC word interpretation (WI) strategies are taught to students to enhance their word- and sentence-level reading ability. Interesting materials and diversified learning activities are used to facilitate students' learning and application of the knowledge and strategies and to enhance their motivation.
- **S**: Teachers teach and model the use of different strategies in the teaching videos. They also check their students' understanding of the strategies, provide constructive feedbacks to improve their strategy use, and facilitate their active participation in the in-class activities.

Flipped Classroom Design:

- **Pre-class e-learning:** Students watch online teaching videos to learn WI strategies and browse the web materials to learn the concept of filial piety to prepare for in-class discussions.
- **In-class learning activities:** Teachers follow up and extend students' pre-class learning. Simple interactive activities are used to facilitate students' application of the learned knowledge and strategies in translating the CC words and achieving basic understanding of the CC texts.
- **Post-class e-learning:** Students practice the WI strategies in an online game platform.

3.2.2 Phase 2: Emulation Level

SRL Instructional Principles: Focus on building up students' foundation on CC reading and emphasis the supportive role of the teacher.

- **T**: All learning materials are selected based on a humanistic theme of learning attitudes. Different types of text comprehension (TC) strategies are taught to students to enhance their text-level reading ability. Interesting materials and diversified learning activities are used to facilitate students' learning and application of the knowledge and strategies and to enhance their motivation.
- **S**: Teachers teach and model how to use different TC strategies in the teaching videos. They also check their students' understanding of the strategies, provide constructive feedbacks to improve their strategy use, and facilitate their active participation in the in-class activities.

Flipped Classroom Design:

- **Pre-class e-learning:** Students watch teaching videos to learn the TC strategies and browse the web materials to understand the attitudes of learning posited by the Confucian scholars.
- **In-class learning activities:** Teacher follow up and extend students' pre-class learning of TC strategies. Interactive and high-level activities are used to facilitate their application of the learned knowledge and strategies in achieving in-depth and higher-level comprehension of the texts.
- **Post-class e-learning:** Students search real-life examples in web and apply the TC strategies and the cultural knowledge in analyzing and discussing the examples.

3.2.3 Phase 3: Self-control Level

SRL Instructional Principles: Focus on students' high-level CC comprehension ability and gradually shift the responsibility from the teacher to students.

- **T**: All learning materials are selected based on a humanistic theme of morality of teachers. Students integrate all knowledge and strategies learned in the previous phases to independently read the texts. Interesting materials and diversified learning activities are used to facilitate their application of the cultural knowledge and strategies, and to enhance their motivation.
- **S**: Teachers facilitate students' active participation in the in-class activities and give constructive feedbacks to them after the activities. Teacher assistance is only provided to students when needed.
- **A**: Most learning activities are student-led. Students can choose different knowledge, strategies, and ways freely to complete the learning tasks.
- **E**: Open and authentic tasks are used to assess students' high-level comprehension. Self- and peer evaluations are added in this module.

Flipped Classroom Design:

- **Pre-class e-learning:** Students browse the web materials to understand the morality of teachers in traditional Confucian culture and apply all learned knowledge and strategies to complete a pre-class assignment.
- **In-class learning activities:** Interactive and high-level activities are used to facilitate students'

application of the learned knowledge and strategies in achieving in-depth and higher-level comprehension of the texts. Teachers guide students to set goals for this module, monitor and evaluate their learning progress and performance.
- **Post-class e-learning:** Students search real-life examples in web, analyze the examples using the learned strategies and cultural knowledge, upload their assignments to an online discussion forum for peer evaluation, and participate in various online discussions.

3.2.4 Phase 4: Self-regulation Level

SRL Instructional Principles: Focus on students' high-level CC comprehension ability and emphasis the active role of the students.

- **T**: Two CC texts with different perspectives on the Confucian views on the human nature are taught to students. Each student group are required to choose a CC text to discuss further the argument of human nature, make a pre-class video introducing the self-selected CC text, and design interactive in-class activities for other classmates to learn and discuss the text.
- **A**: Besides the first two CC texts, all learning materials are chosen by students and all learning tasks are led by students.
- **E**: Authentic and performance-based assessments are used to evaluate students' performance in the pre-class videos and in-class learning activities. Self- and peer evaluations are integrated into the learning activities.

Flipped Classroom Design:

- **Pre-class e-learning:** Students complete an online planning form to set goals and make concrete plans for the module. Each group is responsible to make a pre-class video and watch other groups' videos to prepare for in-class discussions.
- **In-class learning activities:** Each group leads interactive in-class activities for other classmates to learn and discuss their self-selected text. Teachers provide feedbacks to students and facilitate their consolidation of learning.
- **Post-class e-learning:** Students search real-life examples in web, analyze the examples using the learned strategies and cultural knowledge, upload their assignments to an online discussion forum for peer evaluation, participate in various online discussions, and complete an online self-evaluation form to reflect whether they have achieved all the goals and areas for improvement.

4. Conclusion

This paper proposes to integrate SRL and e-learning based on the RDFC model to form a holistic design of an CC reading intervention program to gradually guide students' development of SRL through the observational, emulation, self-controlled, and self-regulation levels. Unlike previous studies that adopted either in-class instruction or e-learning to enhance students' SRL, the intervention program combines both forms of effective learning to overcome the problem of a lack of student autonomy in traditional CC reading instruction and to provide sufficient teacher support for students when they have not fully developed into self-regulated learners. The intervention program will be implemented among Hong Kong junior secondary students in the academic year of 2021-22 and 2022-23. The effectiveness of the program in enhancing students CC reading will be examined using a quasi-experimental design. Quantitative and qualitative methods, including reading tests, questionnaires, interviews, and classroom observations, will be used to collect data for evaluating the program effectiveness and exploring teachers and students' perceptions on the program. The results should shed light on the possibility of combining SRL and e-learning to achieve greater positive effects in an unexplored area of language learning.

Acknowledgements

This article is part of a research project "Enhancing students' classical Chinese reading comprehension and motivation through self-regulated learning-based instruction: A longitudinal and holistic design" funded by the Hong Kong Research Grants Council (RGC 14608820).

References

Anderton, B. (2006). Using the online course to promote self-regulated learning strategies in pre-service teachers. *Journal of Interactive Online Learning, 5*, 156-177.

Berrett, D. (2012). How “flipping” the classroom can improve the traditional lecture. *The Chronicle of Higher Education, 12*, 1-14.

Blau, I., & Shamir-Inbal, T. (2017). Re-designed flipped learning model in an academic course: The role of co-creation and co-regulation. *Computers & Education*, *115*, 69-81.

Chen Hsieh, J. S., Wu, W. C. V., & Marek, M. W. (2017). Using the flipped classroom to enhance EFL learning. *Computer Assisted Language Learning*, *30*(1-2), 1-21.

Chi, L. C., & Chiou, G. F. (2015). Wenyanwen yuedu lijie lichen tanjiu [The comprehension process of reading Classic Chinese texts]. *Huayuwen jiaoxue yanjiu, 12(2)*, 51-74.

Fulton, K. (2012). Upside down and inside out: Flip your classroom to improve student learning. *Learning & Leading with Technology, 39*, 12–17.

Housand, A., & Reis, S. M. (2008). Self-regulated learning in reading: Gifted pedagogy and instructional settings. *Journal of Advanced Academics, 20,* 108-136.

Hsieh, J. S. C., Huang, Y. M., & Wu, W. C. V. (2017). Technological acceptance of LINE in flipped EFL oral training. *Computers in Human Behavior*, *70*, 178-190.

Johnson, G., & Davies, S. (2014). Self-regulated learning in digital environments: Theory, research, praxis. *British Journal of Research*, *1*, 1-14.

Lau, K. L. (2011). Collaborating with front-line teachers to incorporate self-regulated learning in Chinese language classes. *Educational Research and Evaluation, 17,* 47-66.

Lau, K. L. (2013). Chinese language teachers’ perception and implementation of self-regulated learning-based instruction. *Teaching and Teacher Education, 31,* 56-66.

Lau, K. L. (2017). Classical Chinese reading instruction: Current practices and their relationship with students’ strategy use and reading motivation. *Teaching and Teacher Education, 64,* 175-186.

Lau, K. L. (2018). Language skills in classical Chinese text comprehension. *Journal of Psycholinguistic Research, 47*, 139–157.

Lau, K. L. (2020). The effectiveness of self-regulated learning instruction on students’ classical Chinese reading comprehension and motivation. *Reading and Writing, 33,* 2001-2027.

Lee, S. W. Y., & Tsai, C. C. (2011). Students’ perceptions of collaboration, self-regulated learning, and information seeking in the context of internet-based learning and traditional learning. *Computers in human behavior*, *27*, 905-914.

Lo, C., & Hew, K. (2017). A critical review of flipped classroom challenges in K-12 education: Possible solutions and recommendations for future research. *Research and Practice in Technology Enhanced Learning, 12*(1), 1-22.

Lombaerts, K., Engels, N., & van Braak, J. (2009). Determinants to teachers’ recognitions of self-regulated learning practices in elementary education. The Journal of Educational Research, 102, 163-173.

Narciss, S., Proske, A., & Koerndle, H. (2007). Promoting self-regulated learning in web-based learning environments. *Computers in human behavior*, *23*, 1126-1144.

Perry, N. E. (1998). Young children’s self-regulated learning and contexts that support it. *Journal of Educational Psychology, 90,* 715-729.

Schunk, D. H. & Greene, J. A. (2018). Historical, contemporary, and future perspectives on self-regulated learning and performance. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance, 2nd ed*. (pp. 1-15). New York: Routledge.

Schunk, D. H., & Zimmerman, B. J. (1997). Social origins of self-regulatory competence. *Educational psychologist*, *32*(4), 195-208.

Sletten, S. R. (2017). Investigating flipped learning: Student self-regulated learning, perceptions, and achievement in an introductory biology course. *Journal of Science Education and Technology*, *26*, 347-358.

Stoeger, H., Sontag, C., & Ziegler, A. (2014). Impact of a teacher-led intervention on preference for self-regulated learning, finding main ideas in expository texts, and reading comprehension. *Journal of Educational Psychology, 106,* 799-814.

Tseng, M. F., Lin, C. H., & Chen, H. (2018). An immersive flipped classroom for learning Mandarin Chinese: Design, implementation, and outcomes. *Computer Assisted Language Learning*, *31*(7), 714-733.

Turan, Z., & Akdag-Cimen, B. (2019). Flipped classroom in English language teaching: A systematic review. *Computer Assisted Language Learning,* DOI: 10.1080/09588221.2019.1584117.

van Grinsven, L., & Tillema, H. (2006). Learning opportunities to support student self-regulation: Comparing different instructional formats. *Educational Research*, *48*, 77-91.

Wang, J., An, N., & Wright, C. (2018). Enhancing beginner learners' oral proficiency in a flipped Chinese foreign language classroom. *Computer Assisted Language Learning*, *31*(5-6), 490-521.

Zimmerman, B. J. (2000). Attainment of self-regulation: A social cognitive perspective. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13-39). Orlando, FL: Academic Press.

A Flipped Model of Active Reading Using a Learning Analytics-enhanced E-book Platform

Yuko TOYOKAWA[a*], Rwitajit MAJUMDAR[b], Louis LECAILLIEZ[a] & Hiroaki OGATA[b]
[a]*Graduate School of Informatics, Kyoto University, Japan*
[b]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*yflamenca18@gmail.com

Abstract: Flipped learning is an effective learning method embraced by many teachers. However, it is difficult to observe how students are actually engaging in their learning outside of classrooms. By using an e-book reader with an analysis tool called BookRoll (BR), learning attempts can be visualized. This is an ongoing experiment in an active reading (AR) class to measure the effect of Survey, Question, Read, Record, Recite and Review (SQ4R), a common AR strategy in a university in Japan. Sixteen freshmen do their AR assignments through BR, and the classroom activities are organized based on the analysis of their outcomes. The following two research questions are stated: 1) what is the engagement of the students in different phase of the learning activity? and 2) is there any relation between the students' test scores and their flipped active reading performance? The results show that there are relations between students' test scores and their flipped reading performance. Limitations and future implications for data informed teaching are discussed.

Keywords: SQ4R active reading, flipped learning, learning analytics, log data

1. Introduction

Reading is a complicated individual task, which is performed in an individual's mind. Because of that, it is always challenging for teachers to design classroom activities for reading. In a reading class, the first objective might be to improve students' reading comprehension skills. But at the same time, it is important to lead students' individual autonomy so that they can work on reading by themselves outside their class. Flipped learning is one of the popular approaches, which has been called attention to by many educational stakeholders. Through flipped learning, students study concepts of the topic at home in advance so that they can engage in classroom activities actively and cooperatively with other students. However, it has been difficult to gauge how and how many students are working on their assignment outside of their class. This study examines how university freshmen in an AR basic class learn through a common active reading strategy, SQ4R (Survey, Question, Read, Record, Recite and Review) (Wong, 2009; Khusniyah & Lustyantie, 2017). They are asked to conduct a reading assignment as flipped learning by using a learning analytics enhanced e-book reader, BookRoll (BR) (Ogata et al., 2015). The classroom activities are designed based on their assignment outcomes. Two research questions are as follows: 1) what is the engagement of the students in different phase of the learning activity? and 2) is there any relation between the students' test scores and their flipped active reading performance?

2. Literature Review

2.1 SQ4R AR Strategy

Introducing appropriate reading strategies would help students to acquire reading skills and, hopefully, would eventually lead to their motivation and autonomy toward reading. Learning a reading strategy might be essential, especially for learners with lower reading competency. Novice readers may benefit from learning how to read actively and critically.

SQ3R (Survey, Question, Read, Recite and Review) was originally developed by Robinson for acquiring fundamental reading skills and its technique has been favorably applied in many reading classrooms (Robinson, 1946). SQ4R (Survey, Question, Read, Record, Recite and Review) is an extended version of SQ3R (Wong, 2009; Khusniyah & Lustyantie, 2017). Its effectiveness has been examined in many previous studies and the results were preferable (Basar & Gürbüz, 2017; Khusniyah & Lustyantie, 2017). Consequently, implementation of the technique in English classrooms has been suggested. Each phase of SQ4R is listed as follows:

- Survey: Skim and scan the overview of the upcoming reading to visualize the contents before reading.
- Question: Ask questions to oneself from what one has grasped from the Survey phase, and/or associate new information with one's previous knowledge to pay more attention to the reading.
- Read: Engage in reading by keeping questions in mind.
- Record: Take notes or annotations, and use markers for later reflection.
- Recite: Recite the contents in one's own words and write summaries for further retention.
- Review: Reinforce what has been read by reading again, answering questions, and/or going through other exercises.

A study was conducted to observe students' reading comprehension and teachers' perception toward SQ4R strategy in an English reading comprehension class by using the Zoom application (Khusniyah, 2020). It indicates that students improved reading comprehension skills and the teachers showed positive perceptions toward SQ4R, but they also indicated that SQ4R was too systematic and required complicated class-activity preparation. Overall, it was concluded that the application of SQ4R in online learning was recommended to use with media according to the needs of the objectives.

2.2 Flipped Learning

Flipped learning is one of the emergent learning methods in which learners go over the concepts of the topic before participating in the class activities through watching video lectures or studying online in advance (Bergmann & Sams, 2012). Learners come to class with their previous knowledge so that they can actively engage in classroom activities. The previous studies about the effect of flipped learning (Acarol, 2019) highlights that it has positive effects for student achievement, classroom participation and motivation, and students' attitudes.

2.3 Learning Analytics of E-book Based Activities

Learning analytics (LA) is a relatively new emergent study field which focuses on how technologically obtained data can be used to improve teaching and learning with minimum time delay between the capture and the use of data (Elias, 2011). An experiment was conducted to explore the relationship between learning engagement, behavior and achievement of senior high school students by using BR (Chen, 2020). The results from Chen's experiment indicated that the LA e-learning approach could assist not only highly engaged learners, but also improve the achievements of learners with low- and medium-level engagement.

3. Orchestrating a Flipped Model of AR Using BR

BR is a learning analytics enhanced e-book browsing tool, which allows students to browse digital learning materials anytime and anywhere (Ogata et al., 2015). Learning materials can be uploaded in the form of a PDF, and users' logs are recorded and visualized using the analysis tool. BR was used as a main learning platform for flipped reading assignments and some class activities in an online class. It provides features such as a recommender system, a yellow and a red marker, a memo board, and a dictionary function named DicoDico (Lecailliez et al., 2020). Learners can look up unknown words by using DicoDico, and these words can be visualized in the analysis tool. A yellow and a red marker can be used to highlight unknown words, as well as main or important ideas. The memo can be used to leave annotations. A pre- and a post-quiz created by the recommender system can be used to examine the learning gain of the students.

3.1 *Students' Workflow with BR Affordances*

A set of the first 4 phases of SQ4R (Survey, Question, Read and Record) is conducted as a flipped reading assignment over BR. The online synchronous class then follows all the SQ4R steps. The basic flow of the AR activities in both the flipped and online mode is illustrated in Figure 1.

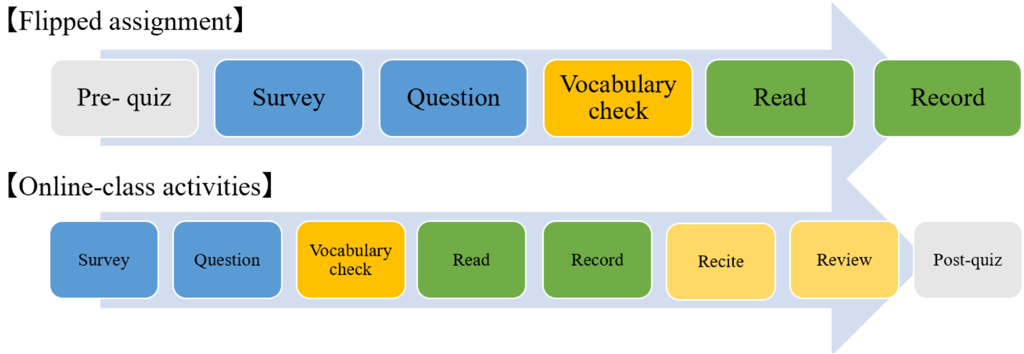


*Figure 1. SQ*4R Flipped Assignment and Online-Class Activities.

The online-class activities are based on what is left in the logs from students' flipped assignment attempts and some requirements indicated by school. As an introduction phase, students' predictions from the Survey and questions are shared, followed by a vocabulary check. Vocabularies highlighted by the students are shared and checked together in class. After confirming the vocabulary and overview of the reading material, students are asked to read the text on BR by themselves. The timer is set on the screen; students can check their reading time on the screen to keep record. After that, students are asked to use markers, DicoDico and the memo board to read the passages carefully during the Record phase. While using Zoom, they can be divided into Zoom breakout rooms and go through the text on BR as a group and confirm the vocabulary and contents together (Read and Recite phases). Using the analysis tool dashboard is recommended to confirm and compare their work with other students' work. After discussing the meaning of the contents, each group answers comprehension questions from the textbook on BR to reinforce their understanding further (Recite phase). Predictions and questions from students' flipped assignment are used as discussion topics for group activities or presentations for further retention of the information. As a part of the Recite and Review phases, they are asked to write a summary of the reading, using the memo board, in Japanese or English, as well as to check their reading speed, review, and take a post-quiz.

3.2 *Teacher's Workflow Over Flipped Active Reading with LA Model*

The teacher's role for the flipped assignments is to upload quizzes and learning materials on BR, provide feedback to students, and assemble online-class activities based on their assignment outcomes. For the online classes, the teacher's role is to assist students as a facilitator: encouraging students to participate in activities, facilitating and monitoring breakout-room activities (if conducted on Zoom) by visiting each room and checking students' engagement with the analysis tool available in near real-time, and dealing with questions from students. A report was distributed as feedback to students and the school based on the data from logs and the results of the first test. Feedback was exchanged between teachers and the school. After reflection, some modifications on flipped assignments and classroom activities were made. The flow of the LA cycle of flipped active reading is illustrated in Figure 2.

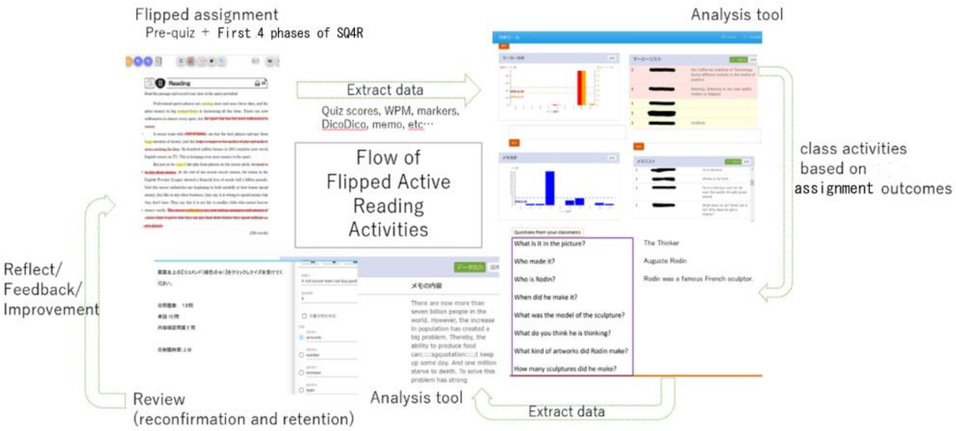

Flipped assignment
Pre-quiz + First 4 phases of SQ4R
Analysis tool
Extract data
Quiz scores, WPM, markers, DicoDico, memo, etc…
Flow of Flipped Active Reading Activities
class activities based on assignment outcomes
Reflect/ Feedback/ Improvement
Review (reconfirmation and retention)
Analysis tool
Extract data

Figure 2. LA Cycle of Flipped Active Reading.

4. Pilot Study

4.1 Study Context and Methods

Participants are 16 university freshmen who are enrolled in an AR basic course. There are three basic classes for the course. Those were divided based on the university enrollment placement test scores. The participants' level is the lowest among the three. The course objectives stipulated by the university are that students will be able to 1) understand simple passages which is about 200 to 250 words, and 2) summarize the reading contents in their own words, and discuss opinions and ideas about the topic. The experimental class is collaboratively implemented by two teachers: one is done by online and the other one is done face-to-face. The AR course lasts one semester, which is a total of 30 classes held twice a week. A class lasts for 90 minutes. For this study, students' engagement during the first 5 flipped assignments and online-class activities were examined.

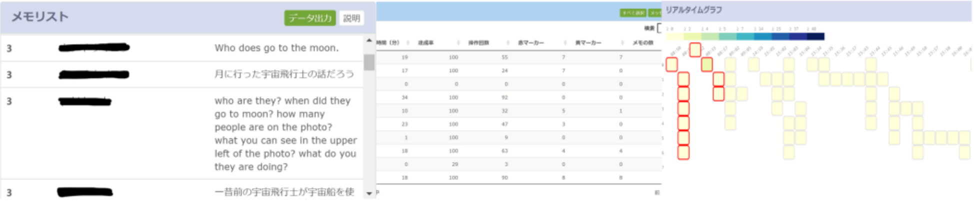
4.2 Data Analysis and Result

Collected data was analyzed with data collected during three different phases: pre- and post-quiz, flipped assignment outputs, and online class activities. One chapter is covered in one class, and a pre- and a post-quiz are assigned for each chapter. The pre-quiz had 15 multiple-choice vocabulary questions, and the post-quiz had 10 vocabularies and 2 comprehension questions to observe students' lexical and reading comprehension achievement. Normalized gain was measured to evaluate the differences between the pre- and post-quizzes in terms of students' performances. 10 out of 16 students attempted to take both quizzes. The average normalized gain score for the experimental group was 21% which was in the low range as shown in Table 1.

Table 1. *Normalized Gain* (as per Hake, 1999)

Normalized gain score(N=10)	Criteria	Value
Average gain		21%
.07<(g)<1.00	High	1 student
0.30<(g)<0.70	Average	4 students
0.00<(g)<0.30	Low	2 students
(g)=0.00	Stable	0 student
minus1.00<(g)<0.00	Decrease	3 students

Extracted data are accumulated into logs in an analysis dashboard. Log data from BR contains students' reading interactions and engagements during the flipped assignments. Memo lists containing predictions and questions were examined for students' comprehension. Students' learning engagements, such as time spent on flipped assignments and the number of operations, were extracted from the feedback panel and real-time graph (see Figure 3).



メモリスト

データ出力 説明

3	[redacted]	Who does go to the moon.
3	[redacted]	月に行った宇宙飛行士の話だろう
3	[redacted]	who are they? when did they go to moon? how many people are on the photo? what you can see in the upper left of the photo? what do you they are doing?
3	[redacted]	一昔前の宇宙飛行士が宇宙船を使

時間（分）	達成率	操作回数	赤マーカー	黄マーカー	メモの数
19	100	55	7	7	
17	100	24	7	0	
0	0	0	0	0	
34	100	92	0	0	
10	100	32	5	1	
23	100	47	3	0	
1	100	9	0	0	
18	100	63	4	4	
0	29	3	0	0	
18	100	90	8	8	

Memo list of flipped assignment Feedback panel for interaction counts Real time graph of flipped assignment

Figure 3. Visualized Data in the Analysis Tool.

Students' predictions and questions left in the memo were examined. Each student wrote a single prediction per each chapter. 7 students wrote all 5 predictions for 5 chapters, while 1 student did not write any prediction as assignment. Regarding questions, a total of 56 questions in English and 24 questions in Japanese were left in the memo.

The operation time as seen in the feedback panel indicated that students were actually using tools such as the markers and the memo while they were studying. It shows that they were not just opening BR, but browsing the pages and actually spending time on studying. The maximum length of the time spent on browsing BR was 75 minutes per day and the average was 20.56 minutes. From the panel, it was understood that 12 out of 16 students did all their flipped assignments for each chapter while 4 students sometimes did not do their assighment. Further, students are required to take four monthly tests and a final exam to fulfill the course grade requirement. Their test results are available from school. For this study, the first monthly test was taken into account. The test assessed students' lexical, grammatical and reading comprehension achievements. The first monthly test scores were ordered from the highest to the lowest (Papi & Abdollahzadeh, 2012). 6 students belong to the high group (score over 91); 7 students are in the middle group (scores from 73 to 90); and 2 students belong to the lowest group (score below 73); 1 student was absent. Furthermore, the transition patterns between the pre-quiz score and students' time engagement in flipped assignments, and the first monthly test scores were illustrated by interactive Stratified Attribute Tracking (iSAT) methods in Figure 4 (Majumdar & Iyer, 2016) in order to answer the second research question for the relation between the students' learning achievement and their flipped active reading performance.

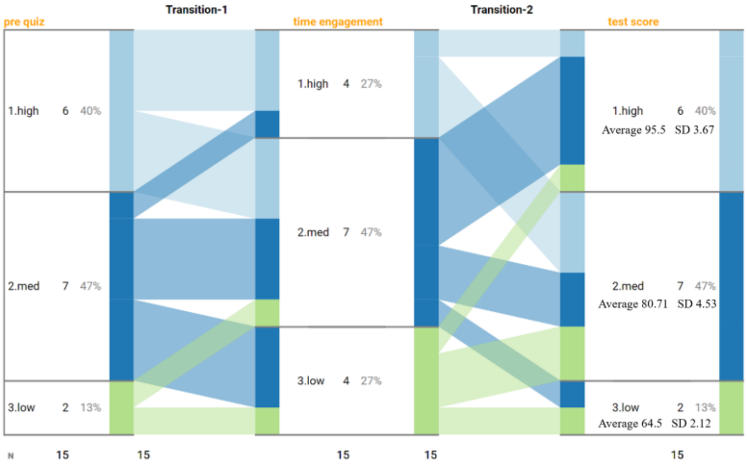


Figure 4. Cohort Transitions between Pre Quiz, Engagement in Flipped Assignments and Term Performance.

It was observed that the test score is related to the quiz score and flipped assignment performance: students who got lower scores on pre-quizzes and spent less time on their homework got lower scores, except one who got a higher score.

5. Discussion and Conclusion

From the implementation of flipped active reading with a learning analytics enhanced e-book reader, students' engagement in each activity phase was examined, and it was observed that there are relations between freshman English learners' learning achievement and their flipped active reading performance. The dashboard also provided the information to the teacher to decide, conduct and monitor the activities during the synchronous online phase.

However, we acknowledge some limitations at this current stage. Overall, the number of participants were less to conduct any inferential statistics related to students' learning performance with SQ4R AR strategy through flipped learning and their achievement. We shall continue with a two-group study design with the current participants as experimental group and other class conducted with traditional activity learning strategy as control groups in order to determine the effectiveness of SQ4R active reading strategy in the flipped learning context.

It took about 4 weeks for most of the students to get used to the learning procedure over digital devices. Flipped learning procedure has to be simple enough so that students are able to attempt learning by themselves. Moreover, giving constant feedback and assistance, and encouraging students to work on their assignments would be the key to conduct flipped active reading effectively.

As for teachers' workload, it takes an extra effort to prepare flipped learning materials and construct class activities based on the students' assignment submission. In addition, technical problems, such as system downtime and bugs, need to be considered and alternatives set aside. The teacher's role as a facilitator and the degree of class preparation has a strong influence on the engagement of the students in the learning activities. Progress of this research aims to further investigate these matters and refine the workflow and technical affordances of the system to support flipped active reading activities.

Acknowledgements

This research was supported by the following grants: JSPS 20K20131, JSPS 16H06304, SPIRITS 2020 of Kyoto University, NEDO JPNP18013, and NEDO JPNP20006.

References

Acarol, K. (2019). A study on the effectiveness of flipped learning model. *Kara Harp Okulu Bilim Dergisi*, *29*(2), 267-295.

Basar, M., & Gürbüz, M. (2017). Effect of the SQ4R Technique on the Reading Comprehension of Elementary School 4th Grade Elementary School Students. *International Journal of Instruction*, *10*(2), 131-144.

Bergmann, J., & Sams, A. (2012). *Flip your classroom: Reach every student in every class every day*. International society for technology in education.

Chen, M. A., Majumdar, R., Hwang, G., Lin, Y. D., Ogata, H., Akcapçnar, G., & Flanagan, B. (2020). Improving EFL students' learning achievements and behaviors using a learning analytics-based e-book system. *In procs. of the 28th International Conference on Computers in Education (ICCE2020)*, Vol.1, pp.474-483.

Elias, T. (2011). Learning analytics. *Learning*, 1-22.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American journal of Physics*, *66*(1), 64-74.

Khusniyah, N. L. (2020). Teacher's Perception on SQ4R in English Reading Comprehension Learning Using Zoom Application. *VELES Voices of English Language Education Society*, *4*(2), 231-238.

Khusniyah, N. L., & Lustyantie, N. (2017). Improving English Reading Comprehension Ability through Survey, Questions, Read, Record, Recite, Review Strategy (SQ4R). *English language teaching*, *10*(12), 202-211.

Lecailliez, L., Flanagan, B., Chen, M.R.A., & Ogata, H. (2020). "Smart dictionary for e-book reading analytics," *In procs. of the 10th International Conference on Learning Analytics & Knowledge*, pp. 89-93.

Majumdar, R., & Iyer, S. (2016). iSAT: a visual learning analytics tool for instructors. Research and practice in technology enhanced learning, 11(1), 16.

Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015). E-Book-based learning analytics in university education. *In procs. of the 23rd International Conference on Computer in Education (ICCE2015)*, pp. 401-406.

Papi, M., & Abdollahzadeh, E. (2012). Teacher motivational practice, student motivation, and possible L2 selves: An examination in the Iranian EFL context. *Language learning*, *62*(2), 571-594.

Robinson, F. P. (1946). Effective study, Rev.

Wong, L. (2009). *Essential Study Skills Sixth Edition*. New York: Houghton Mifflin Company.

Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Modelling the Relationship between English Language Learners' Academic Hardiness and Their Online Learning Engagement during the COVID-19 Pandemic

Lin LUAN[ab]*, Yanqing YI[a] & Jinjin LIU[a]
[a]*School of Humanities, Beijing University of Posts and Telecommunications, China*
[b]*School of Educational Technology, Beijing Normal University, China*
*luanlin@bupt.edu.cn

Abstract: This study presents a structural relationship model that integrates English language learners' academic hardiness with their online learning engagement. Two questionnaires, Academic Hardiness (OH) and Online English Learning Engagement (OLLE), were developed and administered to 453 Chinese university students. The results indicated that AH is composed of four factors, namely commitment, control-effort, control-affect and challenge. Meanwhile, OLLE consists of four factors: behavioral engagement, cognitive engagement, emotional engagement and social engagement. The path analysis revealed that academic hardiness played a positive role in the different aspects of their online learning engagement. Surprisingly none of the sub-dimensions of academic hardiness could predict emotional engagement. Related pedagogical implications are also discussed.

Keywords: Academic hardiness; online learning engagement; EFL learner; COVID-19; structural equation modelling (SEM)

1. Introduction

The outbreak of COVID-19 pandemic has created profound impact on language education, which has traditionally relied on face-to-face instruction (Zhang et al., 2021). During the COVID-19 pandemic lockdown, the Chinese government has issued a policy known as *Suspending Classes without Stopping Learning* in order to ensure that teaching and learning will continue without disruptions. Despite the great affordance provided by the online courses, some problems and challenges were also triggered the emergency conversion to online language teaching, such as distraction from learning tasks, and superficial interactions (Li, 2021). Therefore, it is vital to optimize online EFL learning during this pandemic lockdown (Luan et al., 2020).

2. Literature Review

2.1 Academic Hardiness

According to Evangelia and Spiridon (2016), hardiness is a positive mentality that combines control of affect, control of effort, commitment, and challenge. Specifically, control of affect represents learners' ability to regulate their emotions when encountered with success, stress, and even academic challenges. Control of effort refers to learners' ability to recognize and activate their behavior to work hard and overcome their academic difficulties. Commitment is indicative of students' willingness to put forth sustained effort and make sacrifices to excel academically. Challenge is defined as students' intent to seek out difficult course work and to view these challenges as experiences that will ultimately contribute to their personal growth of academic performance and achievement (Kevin et al., 2019). In the EFL setting, academic hardiness plays an important role for outstanding performance and commitment to goals (Lee, 2020).

2.2 Online English Learning Engagement

In the context of online learning, student engagement refers to the time and energy paid by the students in the process of online learning (Ma et al., 2015). Philp and Duchesne (2016) delineated English learning engagement as multidimensional constructs that influenced by emotion, behavior, cognition of individuals and social factors. Behavioral engagement exhibits their specific learning behavior in the autonomous learning, such as whether they read course resources, answer questions, and complete assignments on time. Emotional engagement refers to their experience towards the learning process and outcome, which covers positive feelings (such as passion, happiness, and enthusiasm) and negative feelings (such as anxiety, burnout and boredom) together. Cognitive engagement includes students' use of learning and self-regulated strategies. In the current research, the four-component model proposed by Philp and Duchesne (2016) is adopted to represent online English learning engagement, consisting of cognitive engagement, behavioral engagement, emotional engagement, and social engagement.

2.3 Academic Hardiness and Online Learning Engagement

Strong associations have been found between academic hardiness and learning engagement. For example, Hodge, Wright and Bonnette (2018) revealed that there is a positive relationship between hardiness and engagement. Besides, Katherine et al. (2017) also indicated that academic hardiness is positively associated with behavioral engagement among college school students. Despite a substantial body of research on L2 learners' hardiness and their learning engagement, their specific structural relations still remain inconclusive, particularly considering these two constructs in an online learning environment. Therefore, this study aims to investigate the intricate interplay between online L2 learners' academic hardiness and learning engagement.

2.4 Research Questions and Hypotheses

This study aims to explore the following two research questions:

1) What are the factorial structure of English language learners' academic hardiness and their online learning engagement?

2) What are the structural relations among the factors of English language learners' academic hardiness and online learning engagement?

According to Katherine et al. (2017), hardiness is an intrinsic psychological quality, whereas engagement belongs to the external behavioral types in the autonomous learning, thus academic hardiness and self-management were the prerequisites for the active social interaction, positive emotion, and the development of individual's skills. Based on this finding, this research regarded academic hardiness as the explanatory variable, while online learning engagement is taken as the criterion. First of all, it is assumed that commitment and control of effort may positively predict their online learning engagement. Then it is proposed that control of affect can also positively predict online learning engagement. On the contrary, it is hypothesized that challenge may negatively explain learners' disengaging in the learning process. As shown in Fig.1, the hypothesized relations among the factors of AH and OLE are presented with the dark and dotted lines indicating the positive and negative relations.

3. Methodology

3.1 Research Context

The present study was conducted in an English course at the first author's university during the semester of the academic year of 2020-2021. During the pandemic lockdown, teachers were directed to deliver online teaching through course delivery tools (e.g. Tencent Classroom), video conferencing platforms (e.g. Tencent Meeting or Zoom) and other social media (e.g. WeChat). A random sample of 453 (74.4% were males) students were involved, ranged from 18 to 23 years' old.

3.2 Instruments

To meet the purposes of this study, we used two instruments: academic hardiness (AH) and online learning engagement (OLE). All the questionnaire items were presented in students' native language, Chinese, on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). Each dimension consists of three to five items.

The AH survey was based on the instrument developed by Benishek et al. (2005). The survey included four dimensions: commitment (e.g. *I would cut back on my extracurricular activities in order to improve my grades*); control of effort (e.g. *I get help when I am not getting the grades I want in school*); control of affect (e.g. *When I do poorly on a test I can stay calm so that I can learn from my mistakes*); challenge (e.g. *I prefer not to take classes that I know are an "easy A"*). To measure students' online English learning engagement, 16 items were from the revised learning engagement scale developed by Luan et al. (2020). Four factors were included: cognitive engagement (e.g. *I try to connect when I am learning online to things I have learned before*); behavioral engagement (e.g. *I complete my online homework on time*); emotional engagement (e.g. *I enjoy learning new things online*); social engagement (e.g. *I try to work with those who can help me online*).

3.3 Research Procedure

First, this study conceptualized the two main research constructs with clearly-defined factors based on the precious research frameworks, and then proposed a hypothesized structural model concerning the relationship among all factors. Then, a structural equation modeling approach has been adopted to test the hypothetical model through confirmatory factor analysis (CFA) and path analysis. Finally, the complex inter-relations among all the factors of the two constructs were investigated. The SPSS 22.0 and AMOS 22.0 were employed to conduct the validity and reliability tests of the two instruments.

4. Results

4.1 CFA Analysis of the Academic Hardiness Survey and Online English Learning Engagement Survey

In order to verify the construct of the academic hardiness (AH) survey, confirmatory factors analysis was conducted. The results showed that all factor loadings were higher than the cut-off value of 0.50. All Average Variance Extracted values (AVE) had exceeded 0.60. The Composite Reliability values (CR) ranged from 0.86 to 0.92. Moreover, all alpha values were above 0.7 and the overall Cronbach's value was 0.92. Therefore, the reliability of the questionnaire was established. In addition, its fit statistics were as follows: χ^2/df =1.75, RMR=0.49, GFI=0.91, NFI=0.92, IFI=0.97, CFI=0.97, based on the Chi-square criterion and the fitting statistics of structural equation model, this survey had good structural validity.

Similar research method was applied to the measurement of online English learning engagement (OELE) survey. The results showed that all Average Variance Extracted values (AVE) of components of OELE had exceeded 0.60, the Composite Reliability values (CR) ranged from 0.86 to 0.90, all alpha values were above 0.7 and the overall Cronbach's value was 0.91. Moreover, $\chi2$/df =1.55, RMSEA=0.50, GFI=0.91, NFI=0.93, IFI=0.98, CFI=0.98. Statistics all indicated that OELE survey had a good reliability and structural validity.

4.2 Descriptive Analysis and Correlation Analysis of Academic Hardiness and Online English Learning Engagement

According to the descriptive analysis, the square roots of the Average Variance Extracted values (AVE) for all constructs were greater than the correlations between constructs, thus academic hardiness showed great discriminant validity. Results also suggested that there is a significant positive correlation among hardiness and online English learning engagement. The higher the hardiness, the deeper online English learning engagement will be.

4.3 Path Analysis

The path analysis was conducted to explore the relationship between learners' academic hardiness and their online English learning engagement. The final structure model is displayed in Fig. 1.

First of all, the results of the path model testing revealed a good model fit with acceptable fitting indices (χ2/df=1.81; CFI=0.90; TLI=0.89 IFI=0.90; RMSEA=0.06). Then a summary of the standardized path coefficients was analyzed, and the associated significance was indicated by asterisks in the figure. As shown in Fig.1, the factor "control-effort" is the most positive factor which significantly predicates three factors of online English learning engagement, with path coefficients ranging from 0.17 to 0.27, all the estimates were statistically significant at $p<0.001$. Learners' challenge can also positively explain the variations in behavioral engagement (β=0.15, $p<0.001$) and social engagement (β=0.23, $p<0.001$). Meanwhile, the factor "commitment" has the positive relationships with cognitive engagement and social engagement. Control-affect is a significant factor for behavioral engagement. Surprisingly, academic hardiness failed to predict learners' emotional engagement, since none of the path coefficients is statistically significant.

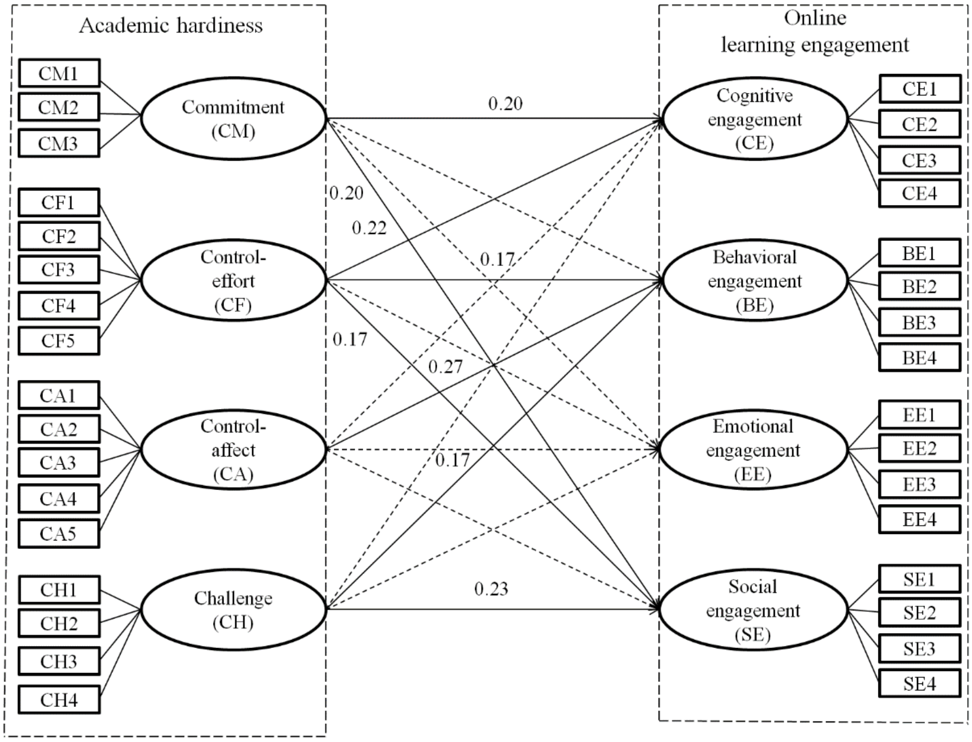


Figure 1. The Final Model of the Structural Relations between the AH and OLE.

5. Discussion and Conclusion

In this study, a proposed model of learners' academic hardiness and learning engagement in English courses was explored in the context of technology-enhanced environment. This research supported the findings of Hodge et al. (2018) in terms of the positive effects of academic hardiness to engagement and academic outcomes for university students.

First, the results suggested that commitment and control-effort may serve as facilitators to students' cognitive engagement. Students may less likely to use deep learning strategies and cope with difficulties if they do not stay focused. Surprisingly, none of the factors of academic hardiness failed to predict learners' emotional engagement. It is also assumed that emotional engagement is more complicated and could be influenced by other factors, such as the learners' technology acceptance and the design of online courses. Second, the results indicated that control-effort, control-affect, and challenge are positive indicators of online behavioral engagement. These results suggest that the more confidence students felt in the learning process, the more sustained effort they put, the more intent to overcoming obstacles, and the more positive attribution they oriented, the more often they would actively engage in the online learning. Third, commitment, control-effort and challenge are significant contributors to online social engagement. Students, who can do correct attributions, strive to overcome temptations and uncertainties in the learning process could maintain their concentration and enthusiasm conditions. Then positive emotions and enough real involvement of students increased the likelihood to interacting with peers and teachers with appropriate approaches.

This study has several limitations that need to be acknowledged. First, the participants of this study were college students majored in science and technology. Future studies should examine the extent to which the current findings would adapt to participants of other demographic characteristics. Second, as all data were collected from participants' self-reported survey, one reasonable step would be to employ multiple methods, such as learning analytics methods using data retrieved from the learning management system (Luan et al., 2020).

Acknowledgements

This research was supported by Teaching Reform Project supported by Beijing University of Posts and Telecommunications (2021KCSZ18), and Teaching Reform Project "High-tech Course" of Intercultural Communication English supported by Beijing University of Posts and Telecommunications.

References

Benishek, L. A., Feldman, J. M., Shipon, R. W., & Mecham, S. D. (2005). Development and evaluation of the revised academic hardiness scale. *Journal of Career Assessment*, 13(1), 59-76.

Hodge & Wright & Bennett. (2018). The role of grit in determining engagement and academic outcomes for university students. *Research in Higher Education*, 59, 448-460.

Katherine, M., Allan, W., & Jiseung Y. (2017). How true is grit? Assessing its relations to high school and college students' personality characteristics, self-regulation, engagement, and achievement[J]. *Journal of Educational Psychology*, 109(5), 599-620.

Kevin, F., Keeley, C., & Shimon, A. S. (2019). How valid is grit in the postsecondary context? A contrast and concurrent validity analysis. *Research in Higher Education*, 60(6), 803-822.

Lee J S. The role of grit and classroom enjoyment in EFL learners' willingness to communicate[J]. *Journal of Multilingual and Multicultural Development*, 2020: 1-17.

Li, B. (2021). Ready for Online? Exploring EFL Teachers' ICT Acceptance and ICT Literacy During COVID-19 in Mainland China. *Journal of Educational Computing Research*, 07356331211028934.

Luan, L., Hong, J. C., Cao, M., Dong, Y., & Hou, X. (2020). Exploring the role of online EFL learners' perceived social support in their learning engagement: a structural equation model. *Interactive Learning Environments*, 1-12.

Ma, J., Han, X., Yang, J., & Cheng, J. (2015). Examining the necessary condition for engagement in an online learning environment based on learning analytics approach: The role of the instructor. *The Internet and Higher Education*, 24,26–34.

Philp, J., & Duchesne, S. (2016). Exploring engagement in tasks in the language classroom. *Annual Review of Applied Linguistics*, *36*, 50-72.

Zhang, C., Yan, X., & Wang, J. (2021). EFL Teachers' Online Assessment Practices During the COVID-19 Pandemic: Changes and Mediating Factors. *The Asia-Pacific Education Researcher*, 1-9.

Design and Evaluation of a Game-based Language Learning Web Application for English Language Learners in Thailand

Kornwipa POONPON*, Wirapong CHANSANAM, Chawin SRISAWAT & Trinwattana POOCHANON[d]
Smart Learning Innovation Research Center, Khon Kaen University, Thailand
*korpul@kku.ac.th

Abstract: This paper proposes a game-based language learning web application, ELA-TIGA, developed for junior high school students to support their English language self-learning. This web application was designed by using a waterfall model and System Development Life Cycle (SDLC) on the Moodle platform and Database Management System (DBMS). The content and games in the ELA-TIGA app were framed by the Task-Input-Genre-Assessment teaching model for language learning. This preliminary study investigated the participating teachers' application evaluation. The participants were sixteen English language teachers from sixteen junior high schools in a north eastern province in Thailand. The teachers were trained how to use the ELA-TIGA app and later tried out the app with their students. After that, they were asked to complete a questionnaire with close-ended and open-ended questions. The findings reveal the teachers' positive feedback about the application’s content and instructional design, user interface, and game and interaction design. Besides, the teachers claimed that a variety of games are easy to use and appropriate for their students' English ability. The application can also motivate their students to learn English. They expected to fully use ELA-TIGA in the near future and believed it would be very useful for their students, especially during an online learning period caused by the COVID-19 pandemic.

Keywords: Technology in education, technology-enhanced language learning, game-based language learning, online games, web application

1. Introduction

In Thailand, technology-enhanced language learning (TELL) has been encouraged to be used by English language teachers to assist their instruction and improve students' English ability (Ministry of Education, 2014). However, it is quite challenging for Thai teachers to use technology in their classrooms as they may not be sufficiently prepared for this pedagogical practice. Many of them may use existing applications such as Kahoot or Flipgrid in their classrooms to help the students learn English with more fun and motivation (e.g., Phenpran & Nahnjun, 2015; Swang phph, 2012). However, the use of these applications may have some limitations. They may neither be in line with learning outcomes in an English course nor related to focused content in students' textbooks. Although the use of the existing applications may help students learn English in class, it may not facilitate their language learning outside the classroom. Being exposed to English anywhere and anytime would accelerate students' success in language learning. It should then be worthwhile to develop an English learning application that can be associated with English course learning outcomes and can promote students' self-learning. This study, therefore, aims to design and develop a web application specially designed for Thai students to learn English actively and meaningfully through game-based language learning (GBLL). It is hoped to fill a gap and shed some light on TELL and GBLL studies in Thailand.

2. Literature Review

The reviewed literature focused on two main concepts used to frame the present study. These involve a review of game-based language learning and online game-based applications used for English language teaching (ELT) in Thai contexts.

2.1 Game-based Language Learning (GBLL)

Game-based language learning (GBLL) is an approach to teaching English where students explore relevant aspects of games in a learning context designed by teachers (Ghazal & Singh, 2016; Taufik, Sabella, & Sabrina, 2020; Wu, Zhang, & Wang, 2020). It is different from gamification in that while gamification is the application of game-like mechanics to non-game contexts to encourage a specific behavior with badges, points, or levels, GBLL involves designing learning activities underpinned by game principles and characteristics and learning theories (Pickles, 2019; Vandercruysse, Vandewaetere, & Clarebout, 2012). For GBLL, games are designed with learning outcomes.

There are several reasons to use games to enhance students' English language ability. First, games provide students with an opportunity to practice vocabulary and grammar, specific functions, and other language skills (Kapp, 2012) as they add variety to the range of learning situations. Games can also increase student-centered learning because they reduce the teacher's dominating roles in the classroom (Willis, 1996). Games not only facilitate students' understanding and development of a second language (Mubaslat, 2012), but they can also engage and motivate language learners (Halleck, Moder, & Damron, 2002; Kapp, 2012; Prensky, 2001; Whitton, 2010). When using games as part of instruction, they can remove boredom without sacrificing the repetition necessary for successful learning (Chitravelu, Sithamparam, & Teh, 1995). Besides offering amusement and cooperation, games are said to help promote positive attitudes towards learning English. They encourage active participation among players and consequently boost confidence and self-esteem. Besides, language games create a relaxing atmosphere. Students are less anxious and more open, and able to communicate when they play games in a language classroom. In sum, introducing games for learners with the intention to teach and further develop their language proficiency is one effective way to enhance language learning.

2.2 Research on Using Online GBLL Applications in Thai Contexts

Online game-based learning has increasingly become popular in ELT. The advantages of games integrated with technology can draw learners into virtual environments that look and feel familiar and relevant. Online or digital games can then meet the needs and learning styles of students in the digital technology era (Prensky, 2001; Wu, Zhang, & Wang, 2020). Thus, it is not surprising to find an emergence of applications for language teaching and learning around the world.

While there are a number of researches on the effectiveness of GBLL applications in foreign contexts (e.g., Alamr, 2019; Taufik, Sabella, & Sabrina, 2020; Wu, Zhang, & Wang, 2020), only a small number of researches in this area have recently emerged in Thailand. Botmart (2019), for example, investigated the use of the application *Classcraft* in teaching vocabulary to Thai university students. The results showed the improvement of students' vocabulary knowledge after using the application and revealed the students' positive attitude towards learning vocabulary through the application. Rachayon (2019) designed digital games for university nursing students to enhance their communicative skills in a flipped environment based on the frameworks of task-based language teaching, digital game-based language learning and flipped learning. She found that not only did the students' oral communication skills improve after learning through the digital games, but the students also had positive attitudes towards the games and perceived the usefulness of the digital games in developing their oral communication ability in English. Although these kinds of studies have supported the advantages of using digital or online games to teach English and improve students' English knowledge and ability, most of them focused on the use of existing applications in ELT. There is still a gap for studies aiming to design, develop, and evaluate ELT applications driven by local English learning problems and teaching contexts.

3. TIGA Model and ELA-TIGA Application

The present study developed an English Language Application: Task-Input-Genre-Assessment (ELA-TIGA) as a web application for English language learning for junior high school students in Thailand. The application aims to enhance students' English ability based on the Task-Input-Genre-Assessment (TIGA) teaching model (Poonpon, Satthamnuwong, & Sameephet, 2016). It is also an extensive game-based learning application that can support students' self-learning. The following describes the TIGA model and how it is used to design the TIGA-based content, followed by the application's architectural and game designs.

3.1 The TIGA Model

The TIGA model was designed under the frameworks of the task-based language learning approach, genre-based approach, the Common European Framework (CEFR), and PISA's reading literacy skills (Poonpon, Satthamnuwong, & Sameephet, 2016). This model was especially developed to address the teaching and learning problems encountered by the teacher and the students in the Thai rural context (Figure 1). It focused on; 1) **Task (T)** for the student to have achievable goals, scaffolding pedagogical tasks, and an authentic target task in local and global contexts; 2) **Input (I)** for students to enhance their vocabulary and grammatical knowledge and use through listening and reading skills needed to complete the tasks; 3) **Genre (G)** as a model for the students to learn communicative functions of a particular language type, and; 4) **Assessment** to help the students evaluate their performance and learn from what they have done in each task. The model was used to design the content of KKU Smart English Books 1 and 2 (Poonpon & Satthamnuwong, 2019; Poonpon, et al., 2019), produced by the Smart Learning Innovation Research Center, Khon Kaen University, Thailand. This content was later adapted to design learning activities or games in the ELA-TIGA application.

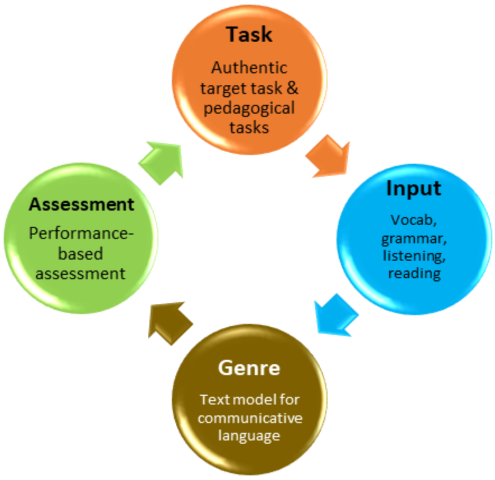

Task
Authentic target task & pedagogical tasks
Input
Vocab, grammar, listening, reading
Genre
Text model for communicative language
Assessment
Performance-based assessment

Figure 1. The TIGA Model (Poonpon, Satthamnuwong, & Sameephet, 2016).

3.2 ELA-TIGA Application' Architectural Design

The ELT-TIGA application was architecturally designed by using a waterfall model and System Development Life Cycle (SDLC). The model is a classic systematical model for developing applications (Pressman, 2005). In this model, each stage fulfills a specific purpose and task. The model includes five stages: studying, planning, creating, testing, and maintenance (Bassil, 2012).

At an initial stage in implementing the ELA-TIGA application, the analysis stage is used to determine teacher and student learning requirements, propose an explanation of an approach in the improvement of eLearning, formulate eLearning's desired ideas, and identify customer requirements. Implementing a web-based and immersive multimedia e-learning framework as features/contents that supplement e-learning is the design stage. Administrators, managers, teachers, and learners are the four main categories of users. Audio (sound effects, background sound, and music), graphics (typography, layout design, colour), creative multimedia (movie, animation), and interactive design are all facets of visual communication that must be addressed when using multimedia (navigation icons).

The open-source software Moodle and Database Management System (DBMS) were used in the development stage of ELA-TIGA construction, allowing the e-learning software to run a server. Using plugins such as H5P, PowerPoint, and other supporting applications, the multimedia capabilities incorporated audio, visuals, video, and animation.

The development of the ELA-TIGA application is defined by integrating interactive multimedia content through different activities, according to H5P (Interactive Content). The teacher/manager will use this exercise to generate content that can enhance the learning paradigm and collaborate to create full interactive content. The teacher would generate interactive multimedia-based learning material with different choices such as image hotspot, image justification, memory games, interactive images, and several more inside this framework by introducing a new functionality called Interactive Content (H5P). As a result, the developed learning media would be more stimulating. For example, several different types of questions can be asked in a quiz, including true/false, multiple-choice, multiple answers, type in, matching, series, numeric, fill in the blanks, multiple-choice, text, and so on. Academy content for a picture, accordion for text, and columns are used to regulate the style of the H5P contents. The learner can organize and apply any form of content to a column as desired.

The architectural design of the ELA-TIGA is shown in Figure 2. In this platform, a person who wants to use eLearning requires a smart device with a web browser and an operating system that supports it. Within the ELA-TIGA application, they are managed by managers and the teachers' role on duty to create and manage H5P contents, eLearning media files, and interactive contents are stored in a cloud service, including a response from the webserver.

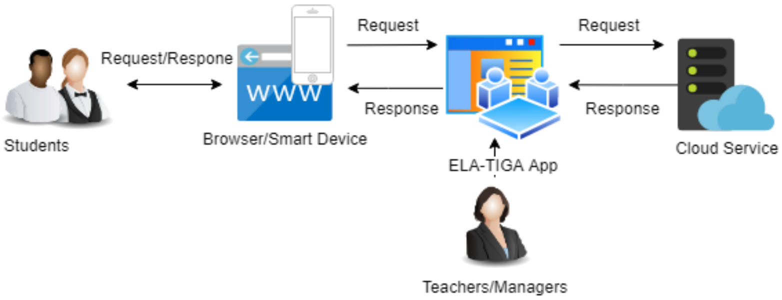


Figure 2. The ELA-TIGA Application’s Architectural Design.

3.3 ELA-TIGA Content and Game Design

The application’s content was adapted from KKU Smart English Books 1 and 2. There are eight units in total. In each unit, input and activities were based on the TIGA model and the four language skills.

All learning activities or games in each unit in the application were designed using the TIGA model. Each game was designed to meet the purpose of each sequence. At the "T" (Task) stage, each unit starts with an introduction to the learning objectives and a target task that students are expected to achieve. This "introduction" part is presented in "Image Hotspot" or "Find the Hotspot" content types.

The second part of the application is "I" (Input), aiming to provide the necessary vocabulary and grammar that are needed for completing the target task. There are around three to four activities under "I," covering vocabulary, grammar, listening, speaking, reading, and writing skills. They are presented in various content types such as Quizzes, Find the Words, Drag and Drop, Memory Games, Dictation, Interactive Video, and Fill in the Blank.

After learning the necessary input, students learn a model text prepared in a specific text type in "G" (Genre) parts. This part involves two activities: reordering parts in the model text and specifying communicative functions of parts of the text. To do this, two content types used in the application are Drag and Drop and Drag the Words.

Finally, "A" (Assessment) aims to assess students' knowledge through reading a model text and answering questions. The questions in this part assessed various sub-skills consisting of accessing and retrieval, integration, and interpretation, and reflection and evaluation skills. Furthermore, there was a self-evaluation form to assess students' understanding of the whole lesson at the end of each unit.

Additionally, a short personality test was also added on the announcement board at the top of the content page. The personality questionnaire, adapted from a personality test developed by Cohen, Oxford, and Chi (2009), aims to reveal students' learning styles and their learning preferences, (i.e., auditory, visual, or tactile/kinesthetic style preferences). Teachers can also use the learning style results when designing materials or building a classroom environment to support students' learning preferences.

4. ELA-TIGA Implementation

The ELA-TIGA application was introduced to sixteen teachers from sixteen junior high schools and two supervisors from the Khon Kaen Provincial Administrative Organization, Khon Kaen, Thailand. They attended a 6-hour training on how to use the application before implementing it with their classes. The implementation was conducted for about one month before a semester ended. After the end of the semester, these teachers completed a five-Likert scale questionnaire and open-ended questions to evaluate the functions of games and give feedback about the overall usefulness of the application.

5. Results and Discussion

The results from the questionnaire show the teachers' demographic information and evaluation of the ELA-TIGA application. All of them hold a bachelor's degree in TESOL or English. About 60% of these teachers have less than 15 years of teaching experience, and the rest have more than 16-years teaching experience. About 70% of them rated themselves to have good technology skills.

Overall, the teachers were satisfied with the ELA-TIGA application. In terms of content and instructional design, all of them agreed that the app content corresponded with the book content, and it was well grouped and put in an appropriate sequence of difficulty. One-third of the teachers moderately agreed that the learning activities or games are neither too difficult nor too easy. They supported that a combination of complex and easy activities would meet the needs of groups of different proficiency levels. In terms of user interface, most teachers believed that the app contains simple components and layout with which their students can use and interact easily. The easy navigation should be practical as this does not require students' to have a great deal of experience. The app is also user-friendly in that it is full of colorful pictures and images, easy to see with appropriate font sizes. When asking about game and interaction design, most of them thought the app contained a variety of games that can encourage students' interaction. They also thought that the application was easily accessible and appropriate for their students' language learning. It is also good to show students their own performance as soon as they finished their activities.

6. Conclusion

The ELA-TIGA application was developed to support students' self-study through game-based language learning along with their Smart English textbooks. In this early stage, the ELA-TIGA application was introduced and somewhat tried out by the English teachers in provincial schools in Thailand. All the teachers' feedback is very important for our team to improve the application's quality. However, this study has limitations. It used a very small sample size and was limited to one group of participants. Future research should focus on the implementation of the ELA-TIGA application with a much larger group of both teachers and students to maximize its usefulness and successfully support Thai students' English language learning.

Acknowledgments

We would like to thank the Smart Learning Innovation Research Center, Khon Kaen University, for funding this app development.

References

Alamr, A. S. (2019). *Digital games and English as Foreign Language (EFL) learning in tertiary education in Saudi Arabia* [Unpublished doctoral dissertation]. University of Wollongong, Australia.

Bassil, Y. (2012). A simulation model for the waterfall software development life cycle. *arXiv preprint arXiv:1205.6904*.

Botmart, V. (2019). *The effects of gamified flipped classroom application on learning English vocabulary for Thai university students in EFL context* [Unpublished master's thesis]. Suranaree University of Technology.

Chitravelu, N., Sithamparam, S., & Teh, S. C. (1995). *ELT methodology: Principles and practice*. Shah Alam: Penerbit Fajar Bakti Sdn. Bhd.

Cohen, A. D., Oxford, R. L., & Chi, J. C. (2009). Learning style survey: Assessing your own learning styles. In B. Kappler Mikk, A. D. Cohen, R. M. Paige, J. C. Chi, J. P. Lassegard, M., Maegher, & S. J. Weaver (Eds.), *Maximizing study abroad: An instructional guide to strategies for language and culture learning and use* (pp. 153-161). CARLA. http:// www.carla.umn.edu/maxsa/documents/LearningStyleSurvey.pdf.

Ghazal, S., & Singh, S. (2016). Game-based language learning: Activities for ESL classes with limited access to technology. *ELT Voices*, *6*(4), 1-8.

Halleck, G. B., Moder, C. L., & Damron, R. (2002). Integrating a conference simulation into an ESL class. *Simulation & Gaming, 33*(3), 330-344.

Kapp, K. M. (2012). *The gamification of learning and instruction: game-based methods and strategies for training and education*. John Wiley & Sons.

Ministry of Education. (2014). *English language teaching and learning reform policy* (pp. 1-20). Office of The Basic Education Commission. http://english.obec.go.th/index.php/download (in Thai)

Mubaslat, M. (2012). The Effect of using educational games on the students' achievement in English language for the primary stage. http://files.eric.ed.gov/fulltext/ED529467.pdf

Phenpran, N. & Nahnjun S. (2015). A study of English vocabulary using the instructional games of Prathom Suksa six students, Wat Thung Noi School, Kui Buri District, Prachuapkhirikhan Province. *Journal of Veridian E-Journal*, *8*(2), 1672-1684. (in Thai)

Pickles, M. (2019). Gamification versus game-based learning. https://www.cambridgeassessment.org.uk/summit-of-education-2019/afternoon-sessions/gamification-versus-game-based-learning/

Poonpon, K., Satthamnuwong, B., & Sameephet, B. (2016). Development of English language teaching model for junior high school students in Northeast of Thailand. Research report. Khon Kaen, Thailand.

Poonpon, K., & Satthamnuwong, B. (2019). *Smart English Student's Book 1* (3rd ed.). Khon Kaen, Thailand: Smart Learning Innovation Research Center.

Poonpon, K., Saengpakdeejit, R., Scherer, A., & Phantharakphong, P. (2019) *Smart English Student's Book 2* (3rd ed.). Khon Kaen, Thailand: Smart Learning Innovation Research Center.

Prensky, M. (2001), "Digital Natives, Digital Immigrants Part 2: Do They Really Think Differently?", *On the Horizon*, *9*(6), pp. 1-6. https://doi.org/10.1108/10748120110424843

Pressman, R. S. (2005). *Software engineering: a practitioner's approach.* Palgrave Macmillan.

Rachayon, S., & Soontornwipast, K. (2019). The effects of task-based instruction using a digital game in a flipped learning environment on English oral communication ability of Thai undergraduate nursing students. *English Language Teaching, 12*(7), 12-32.

Swang phph, C. (2012). Improving English speaking skills using interactive games for students in Prathom 6, Ban Khon Tae Sap, School 3. Sisaket: Ban Khon Tae School. (in Thai)

Taufik, P. H., Sabella, E. N., & Sabrina, S. M. (2020). The use of digital game-based learning in EFL classroom: Teacher's voices. *Proceedings of the International Conference on English Language Teaching (ICONELT 2019).* https://doi.org/10.2991/assehr.k.200427.056

Vandercruysse, S., Vandewaetere, M., & Clarebout, G. (2012). Game-based learning: A review on the effectiveness of educational games. In M. M. Cruz-Cunha (Ed.), *Handbook of research on serious games as educational, business, and research tools* (pp. 628–647). IGI Global

Whitton, N. (2010). *Learning with digital games: A practical guide to engaging students in higher education*. Routledge.

Willis, J. (1996). *A framework for task-based learning*. Pearson PTR.

Wu, Q., Zhang, J., & Wang, C. (2020). The effect of English vocabulary learning with digital games and its influencing factors based on the meta-analysis of 2,160 test samples. *International Journal of Emerging Technologies in Learning, 15*(17), 85-100. http//:academic.obec.go.th/images/document/1525235513_d_1.pdf

The Learning Potential of Online Student-Generated Questions Based on Given Graphics for English Language Learning

Fu-Yun YU
Institute of Education, National Cheng Kung University, Taiwan
fuyun.ncku@gmail.com

Abstract: The student-generated questions approach has received increasing attention over recent decades, and empirical evidence has generally substantiated its positive learning effects. Despite this, existing evidence mostly comes from studies involving students generating questions based on a given text. Since different structures and information specifications are revealed when materials are presented graphically and textually, the learning potential of student-generated questions based on given graphics was the focus of this study. Specifically, both quantitative and qualitative data on the relative learning usefulness of student-generated questions based on given graphics versus text was collected from two classes of sixth-graders (*N*=47) learning English. Five major findings were obtained. First, comparatively, significantly more participants voted for student-generated questions based on given graphics as better supporting their learning of English as compared to student-generated questions based on given text. Second, the X^2 test on the perceived learning usefulness of the two approaches was statistically significant, $X^2 = 8.34$, $p < .001$. Thirdly, the constant comparative method conducted on the explanatory reasons highlighted two learning effects associated with student-generated questions based on given graphics: facilitating the target word to be better understood and memorized, with 'associative links' and 'prompting' noted as the underlying mechanism. Fourth, the learning usefulness associated with student-generated questions based on text focused on its facilitating effects on memorizing the spelling of the target word. Finally, the distinct affordances associated with student-generated questions based on stimulus given in graphics and text forms, respectively, were recognized by those expressing similar perceptions toward the two forms, which tap into different important aspects of the learning of vocabulary in second language learning, namely, pronunciation, spelling, and meaning. Limitations of this study, including order effects, data obtained mainly based subjective evaluation, and vocabulary as the content-to-be-learned are highlighted, and suggestions for future studies are provided.

Keywords: English language learning, graphics as a stimulus, online learning activity, perceived learning potential, student-generated questions

1. Introduction

Currently, the idea of enabling learners to be knowledge producers (rather than knowledge consumers) has been embraced by educators at various educational levels (Persada, Ivanovski, Miraj, Nadlifatin, Mufidah, Chin, & Redi, 2020; Snowball & McKenna, 2017). Studies involving learner production of instructional materials and learning objects (i.e., the learner-generated content approach) have evidenced increasingly wide acceptance of the student populations in different educational contexts (Hubbard, Jones, & Gallardo-Williams, 2019; Orús, Barlés, Belanche, Casaló, Fraj, Gurrea, 2016). Over recent decades, among the various possible forms, student-generated questions have received growing attention from both practitioners and researchers as an efficacious approach for teaching, learning and assessment (Yu, Wu, & Hung, 2014).

Rather than relying on teachers as the question-author as has been conventionally done, the student-generated questions approach accentuates the pedagogical value of having students play the role of the question-author for a benign change (Yu & Wu, 2020). Being entrusted with this empowering role and new task, students are believed to focus on areas of importance and relevance in the learning material as targets of question-generation throughout the learning process (Yu, Wu, & Hung, 2014). In general, empirical evidence has substantiated the positive effects of learner-generated questions in terms of promoting cognitive, affective, and social development (Yu & Wu, 2020). Despite its overwhelming success, an analysis of existing studies reveals that most research on student-generated questions deals with situations where students were directed to generate questions based on a given text (e.g., a specific given question, a reading passage, etc.). The learning potential of question-generation based on a picture, chart, or table (that is, the semi-structured problem-posing situation as classified by Stoyanova and Ellerton) (1996) is yet to be exploited and fully understood.

As suggested and substantiated by researchers in the communication and multimedia field, message conveyed in text and graphics forms tend to elicit different cognitive and emotional reactions in the message-receivers (Clark & Mayer, 2011). Non-verbal information is distinctive in terms of attracting attention, illustrating complex processes, and visualizing abstract content (Burnye, Ditman, Augustyn, & Mahoney, 2009; Lenzner, Schnotz, & Müller, 2013; Wylie & Chi, 2014). In light of their differences, instructional materials usually contain verbal as well as non-verbal content (Ainsworth, 1999).

Taking into consideration that different information structures and specifications will be revealed by stimulus materials presented in the form of graphics and text, as well as the fact that today's learners are born in the digital age and are thus used to a media-rich learning environment (Prensky, 2001), the learning potential of student-generated questions based on given graphics serves as the focus of this study. Specifically, students' perceptions of the learning usefulness of student-generated questions based on given graphics versus text are examined in the present work.

2. Methods

2.1 Participants and Context

A group of students from two sixth-grade classes (N = 47) in a single elementary school were invited to participate in an online student-generated questions activity during their normal 40-minute computer literacy class. The integrated online learning activity was introduced to help enrich student English learning. Particularly, vocabulary was targeted in these sessions in view of its instrumental role in learning a foreign language (El-Nekhely, El-Dien, Al-Hadi, & Khodary, 2019; Folse, 2004), its facilitating effects for the mastery of four language skills (Herman & Dole, 1988), and its importance to school overall academic performance (Blachowicz, Fisher, Ogle, & Watts-Taffe, 2006).

2.2 Online Learning Activity and Study Procedures

An online learning system developed by the research team led by the author (Yu, 2021) was adopted to support the learning activity for a duration of nine weeks. As a routine, after attending four 40-minute instructional sessions on English in a two-week time-frame, the participants headed to the participating school's computer lab for the online learning activity.

This study consisted of three main phases: training (Phase I), student-generated questions based on given text (i.e., a set of vocabulary covered in the current lesson) (Phase II), and student-generated questions based on given graphics (i.e., the same set of vocabulary covered in the current lesson but depicted in a pictorial form) (Phase III). Each of the phases was briefly explained.

During Phase I (i.e., the training session), firstly, an explicit framework devised by the author was introduced, and steps involved in generating questions (in the form of three hints) by referring to this framework for vocabulary of one's choice was explained. As can be seen, the devised framework aims to help the participants focus on the various important aspects of English vocabulary — spelling, pronunciation, meaning, and related words (i.e., at least three words or phrases that would be meaningfully associated with the vocabulary).

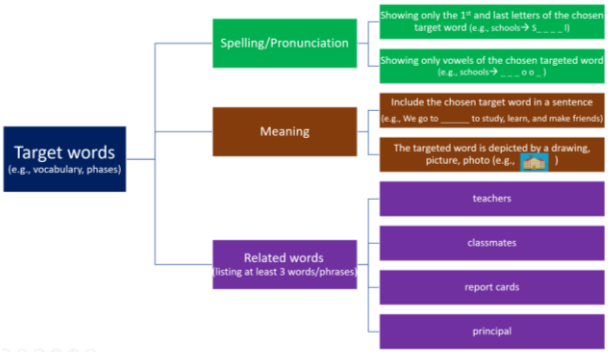

Target words
(e.g., vocabulary, phases)
Spelling/Pronunciation
Showing only the 1st and last letters of the chosen target word (e.g., schools→ S_ _ _ _ l)
Showing only vowels of the chosen targeted word (e.g., schools→ _ _ _ o o _)
Meaning
Include the chosen target word in a sentence (e.g., We go to ______ to study, learn, and make friends)
The targeted word is depicted by a drawing, picture, photo (e.g.,)
Related words
(listing at least 3 words/phrases)
teachers
classmates
report cards
principal

Figure 1. The Framework Devised in Support of Student-generated Questions (in the Form of Hints) for English Learning.

As shown in Figure 1, a set of three hints generated for the vocabulary of ‘school’ could be (1) s_ _ _ _ l or _ _ _oo__; (2) We go to _ _ _ _ _ _ _ to study, learn, and make friends, or a graphics of student’s choice to depict schools; (3) teachers, classmates, report cards, and principal. Another illustrative example generated by one of the study participants for the vocabulary of ‘play basketball’ was:

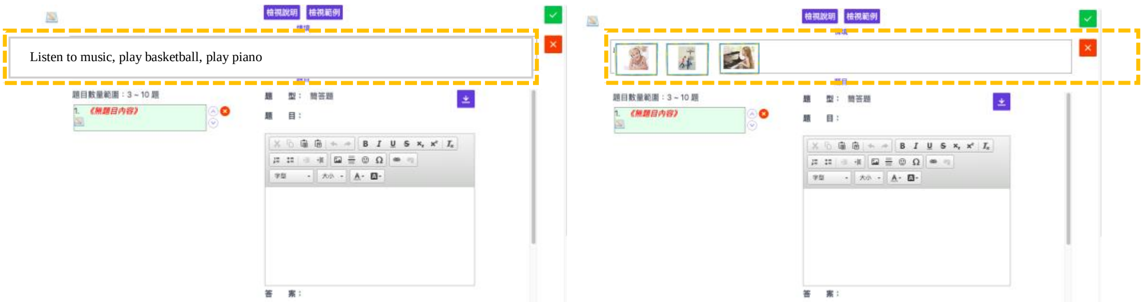
(1) p_ _y b_ _ _ _ _ _ _ _ _l

(2)

(3) NBA, dribble, shoot, Michael Jordan

Afterwards, the criteria for good student-generated questions and the operational procedures involved in navigating the adopted online system were explained. Criteria of good student-generated questions introduced during the training session include: correct spelling, no grammatical errors, and hints helping build meaningful word links to the vocabulary, among others. The training session concluded with a brief practice session on student-generated questions in the system.

During both Phases II and III (a total of four sessions), for each session the participants were directed to generate three questions, each of which consisted of a set of three hints for the vocabulary covered in English class. The only difference is that during Phase II, the vocabulary for question-generation were presented in text form (the left of Figure 2), whereas during Phase III, they were presented in a graphics form (the right of Figure 2). In other words, the participants in this study generated questions with reference to a set of vocabulary, which served as stimuli and were presented in either text or graphics forms.



Listen to music, play basketball, play piano

題目數量範圍：3～10 題

1. 《無題目內容》

題　型：簡答題

題　目：

答　案：

題目數量範圍：3～10 題

1. 《無題目內容》

題　型：簡答題

題　目：

答　案：

Figure 2. Phase II: Online student-generated questions based on given text (left); Phase III: online student-generated questions based on given graphics (right).

After the conclusion of the last online student-generated questions session, one closed-ended question with explanations for one’s selection was distributed to solicit the participants’ views towards the relative learning usefulness of online student-generated questions based on given text and graphics — which of the two question-generation approaches do you think better promote your learning of English vocabulary covered during this study: student-generated questions based on given text, student-generated questions based on given graphics, both at about the same level. Explain your selection.

3. Results and Discussion

As shown in Table 1, nearly half of the participants regarded ‘online student-generated questions based on given graphics’ as better promoting their learning of English vocabulary. Comparatively, significantly fewer participants voted for ‘online student-generated questions based on given text,’ and more than one-third of the participants felt that the two approaches were at about the same level in terms of learning usefulness. Furthermore, an X^2 test on the observed frequency distribution among the three options conducted was statistically significant, $X^2 = 8.34$, $p < .001$.

Table 1. *Descriptive and Inferential Statistics for Student-Perceived Learning Usefulness of Online Student-Generated Questions Based on Given Text and Graphics* (*N*=47)

	SGQ* based on given text *f* (%)	SGQ* based on given graphics *f* (%)	About the same level *f* (%)	X^2	*p*
Learning usefulness	7 (14.89%)	23 (48.94%)	17 (36.17%)	8.34	.01

*SGQ: student-generated questions

The constant comparative method proposed by Lincoln and Guba (1985) was adopted to analyze the descriptive explanations provided by the participants accompanying their selections. Mainly, two processes were involved for the analysis of the descriptive data collected — unitizing and categorizing, for which several steps were involved. First, each and every data entry was read closely and repeatedly to gain a broad sense of the rendered responses. Second, the words, phrases, and/or sentences that captured the key ideas/concepts of individual messages were highlighted (by underlying or circling), and the first impression or thoughts on the data were written on the side of each entry. Third, individual code labels that emerged from the text as the initial coding scheme with definitions were compiled. Fourth, the derived coding scheme underwent several rounds of revisions/refinement, and a final set of categories was set until the resultant categorizations were mutually exclusive, and complete consistency was achieved. The main themes emerged for each of the three options are presented and discussed separately below.

First, for those rooting for student-generated questions based on given graphics, two salient features emerged, and both evolved around learning effects. Specifically, more than half of this group of participants (*N*=13) highlighted the beneficial effects of graphics for helping the vocabulary in target to be ‘better understood’ while more than one-fourth of this group (*N*=6) noted graphics facilitating effects for ‘memorization.’ Although most of the respondents did not elaborate on the underlying mechanism for the perceived learning gains, three made explicit remarks on ‘the associative link’ that graphics offered, and two wrote about graphics prompting them to the question-generation task that they would not otherwise do. The associative link noted by the participants reflected on the ‘referential connections’ put forth in Paivio’s (2001) dual coding theory. Simply put, dual coding theory highlights the essential role of forming and establishment of referential connections between verbal and nonverbal representations for facilitating future retrieval from memory and better comprehension and learning on the part of learners (Clark & Paivio, 1991). The prompting effects noted by the participants resonated well with what multimedia researchers have noted about the ability of non-verbal messages to attract attention (Wylie & Chi, 2014) and elicit emotional reactions (Clark & Mayer, 2011).

As for those supporting student-generated questions based on given text, one major theme emerged. Specifically, more than half of this group of participants (i.e., four out of seven) pinpointed that a given text (acting as a stimulus for online student-generated questions) helped them memorize the spelling of the target word. It should be noted that two participants from this group commented on the explicit nature of the text, which helped eliminate a lack of clarity, which may be the case for graphics. In other words, the definite and explicit nature of the text in terms of directing the learners’ attention to the task (i.e., generating questions for vocabulary) was noted and appreciated by those leaning toward student-generated questions based on text acting as stimuli.

Finally, for those feeling both forms provided similar support for learning, one major theme was revealed, pointing to the benefits provided by the respective forms. Specifically, five of this group of participants revealed that the text form helped the spelling and pronunciation of vocabulary to be better learned while the graphics form helped vocabulary to be ‘easier,’ ‘better understood,’ or remembered due to ‘the situational context revealing within the graphics.’ In other words, the meaning aspect of language learning was supported better by the graphics whereas the spelling and pronunciation aspects of language learning were tapped better by the text.

4. Conclusions

As found in this study, comparatively significantly more participants voted for student-generated questions based on given graphics as better supporting their learning as compared to given text. The explanatory reasons the participants provided shed further light on the distinct affordances associated with student-generated questions based on a given text and graphics. To summarize, as revealed based on the constant comparative method, two learning effects were highlighted by the participants voting for the graphics form: facilitating the target word to be better understood and memorized, with associative links and prompting noted as the underlying mechanism. Alternatively, the learning usefulness of the text form centered on helping students to memorize the spelling of the target word. Finally, the benefits associated with the two different forms were recognized by those expressing similar perceptions towards the two forms, which helped to tap into different important aspects of vocabulary learning in second language learning, namely, pronunciation, spelling, and meaning.

4.1 Limitations of This Study

Although the results of this study revealed some distinctive aspects of online student-generated questions based on given text and graphics, and the participants were found to associate different learning gains related to question-generation based on the two different given stimuli, some limitations of this study should be noted.

First, order effects may be in existence. As described, Phase 2 (student-generated questions based on give text) and Phase 3 (student-generated questions based on graphic) in this study were fixed. It is possible that after exposure to Phase 2, the participants may naturally become more familiar and proficient at the question-generation task during Phase 3, and thus regarded student-generated questions based on graphic as more effective.

Second, data in this study were based on the participants’ subjective rather than objective evaluation.

Third, in this study, English vocabulary learning was the focus. As can be expected, if grammars (i.e., subjects of an abstract nature) were the to-be-learned content, the perceived learning usefulness of text versus graphics may alter. This is especially relevant when noting the respective distinct features of text and picture representations. As pointed out by Schnotz (2014), text is suited for description and explication of abstract concepts whereas graphics are equipped at portraying and depicting concrete objects in a holistic matter (e.g., the size, color, details of the targeted object).

4.2 Suggestions for Future Study

In light of the preliminary nature of this study and the findings obtained, suggestions for future study are offered. First, studies attending to order effect (via conducting Phase II and Phase III alternately), collecting data on learning performance based on objective assessment (e.g., academic achievement, student question-generation performance, etc.), and extending to different focus of language learning (e.g., grammar) would be warranted.

Additionally, issues regarding if there are any differential learning outcomes between the text and graphics forms await further investigation in the future. As noted by the participants in this study, student-generated questions based on given graphics and text may reveal different degrees of information specifications, which may prompt learners to engage at different levels. Therefore, their comparative effects on engagement with a task, learning motivation, academic emotions, and academic performance would be a topic worthy of examination via an experimental research method.

Finally, this study examined the relative learning potential of student-generated questions based on a given text and graphics. Considering that instructional materials usually contain both verbal and non-verbal content for the purpose of better communication while striving to attain complementary or mutually reinforcing functions (Ainsworth, 1999), the effects of student-generated questions based on a given text, graphics, and text alongside graphics on learning would be an interesting extension to be examined in the future.

Acknowledgements

This study was supported by the Ministry of Science and Technology in Taiwan (MOST 108-2511-H-006-007-MY3). Thanks are extended to Zi-Yu Hsu for serving as the research assistant to collect the data reported in this work.

References

Ainsworth, S. (1999). The functions of multiple representations. *Computers & Education, 33*, 131-152.

Blachowicz, C., Fisher, P., Ogle, D., & Watts-Taffe, S. (2006). Theory and research into practice: Vocabulary: Questions from the classroom. *Reading Research Quarterly, 41*(4), 524-539.

Burnye, T. T., Ditman, T., Augustyn, J. S., & Mahoney, JC. R. (2009). Spatial and nonspatial integration in learning and training with multimedia system. Zheng, R. Z (Ed), In *Cognitive effects of multimedia learning* (pp. 108~133). Hershey, PA: Information Science Reference.

Clark, R., & Mayer, R. (2011). *E-learning and the science of instruction*: *Proven guidelines for consumers and designers of multimedia learning* (3rd ed.). San Francisco, CA: John Wiley & Sons.

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review, 3*, 149-210.

El-Nekhely, M. A. E., El-Dien, A. H. S., Al-Hadi, T. M. & Khodary, M. M. (2019). Using pictorial stories for the acquisition and retention of English vocabulary in kindergarten. Retrieved from https://journals.ekb.eg/article_87526_e8d3815c0517edf1074d2e0fbe2d711c.pdf DOI: 10.21608/jfes.2019.87526

Folse, K. (2004). The underestimated importance of vocabulary in the foreign language classroom. *Language Teaching Research, 12*(3), 329–363.

Herman, A. & Dole, J. (1988). Theory and practice in vocabulary learning and instruction. *The Elementary School Journal, 89*(1). https://doi.org/10.1086/461561

Hubbard, B. A., Jones, G. C., & Gallardo-Williams, M. T. (2019). Student-generated digital tutorials in an introductory organic chemistry course. *Journal of Chemical Education, 96*(3), 597-600.

Lenzner, A., Schnotz, W., & Müller, A. (2013). The role of decorative pictures in learning. *Instructional Science, 41*, 811-831.

Lincoln, Y. S. & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage Publications.

Orús, C., Barlés, M. J., Belanche, D., Casaló, L., Fraj, E., Gurrea, R. (2016). The effects of learner-generated videos for YouTube on learning outcomes and satisfaction. *Computers & Education, 95*, 254-269.

Paivio, A. (2001). *Mind and its evolution: A dual-coding approach*. Mahwah, NJ: Erlbaum.

Persada, S. F., Ivanovski, J., Miraj, B. A., Nadlifatin, R., Mufidah, J., Chin, J., & Redi, A. A. N. P. (2020). Investigating generation Z' intention to use learners' generated content for learning activity: A theory of planned behavior approach. *International Journal of Emerging Technologies in Learning, 15*(4), 179-194

Prensky, M. (2001). Digital natives, digital immigrants. *On the Horizon 9*(5), 1-6.

Schnotz, W. (2014). Integrated model of text and picture comprehension. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 72–104). (2nd ed.). Cambridge: Cambridge University Press.

Snowball, J. D., & McKenna, S. (2017). Student-generated content: An approach to harnessing the power of diversity in higher education. *Teaching in Higher Education, 22*(5), 604-618.

Stoyanova, E. & Ellerton, N. F. (1996). A framework for research into students' problem posing in school mathematics. *Proceedings of the 19th Annual Conference of the Mathematics Education Research Group of Australasia* (MERGA), June 30 - July 3, 1996 at the University of Melbourne.

Wylie, R., & Chi, M. T. H. (2014). The self-explanation principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 413–432). Cambridge University Press.

Yu, F. Y. (2021). Development and preliminary evaluation of the learning potential of an online system in support of a student-generated testlets learning activity. *Proceedings of the 29th International Conference on Computers in Education*. November 22-26, Bangkok, Thailand.

Yu, F. Y. & Wu, W. S. (2020). Effects of student-generated feedback corresponding to answers to online student-generated questions on learning: What, why, and how? *Computers & Education, 145*, 103723.

Yu, F. Y., Wu, C. P., & Hung, C-C (2014). Are there any joint effects of online student question generation and cooperative learning? *The Asia-Pacific Education Researcher, 23*(3), 367-378.

Proctored vs Unproctored Online Exams in Language Courses: A Comparative Study

Mehmet Ali ÇELİKBAĞ*& Ömer DELİALİOĞLU
Department of Computer Education & Instructional Technology, Middle East Technical University, Turkey
*celikbag@gmail.com

Abstract: This study investigates online exams of online language courses at a distance education center of a public university in Turkey. Causal comparative research design was followed to examine issues in online language courses with the aim to better develop online exams. Participants were 105 students from spring semester of 2019 and 116 students from spring semester of 2020 from six different programs at associate degree level at a public university. The overall findings indicated that there were varied results in unproctored online exams (UOE) of online Turkish (written and verbal expression) and English (grammar) language courses. Effects of the COVID-19 pandemic were observed in an unproctored online final exam for the English language course. However, tests showed that there was a high reliability when they were administered in different years to similar conditions and groups of students. As an online exam platform, Moodle LMS was used. Students of Computer Programming and Justice programs in online language courses had greater achievement results when compared to other educational programs whether online or paper-based and proctored or unprocotored exams were conducted. Although gender did not play a vital role in achievement results in general, it was highly influential in the unproctored online English final exam in favor of male students.

Keywords: Online exams, unproctored online exams, advantages and disadvantages of online exams, online language learning, distance education, COVID-19 pandemic effects on education

1. Introduction

The proportion of distance education is increasing, and this creates implications for design and implementation of courses by considering new challenges and advantages. Recently, the COVID-19 pandemic has created further implications in distance education, especially in terms of examination processes. This study aims to investigate online exams with the aim of finding ways to improve assessment experience and quality, and to analyze the effects of settings and conditions. Within the scope of this study, quality, reliability and validity issues regarding online summative assessment in distance education are discussed in a holistic way. Specifically, the study seeks to provide insights for effective design of online exams. Distance education and its evaluation for the improvement online exams are investigated with causal-comparative study. Since the beginning of the pandemic many countries have switched to open and distance education settings. In this study, six educational programs at a distance education center of a public university in Turkey were studied.

2. Literature Review

Online language learning is increasingly being welcomed in formal and informal education environments. Since Open University's online French course in 1995 and the last decade of 20th century, many developments have taken place such as learning languages online through MOOCs, mobile applications in online learning, especially in informal settings, and virtual worlds (Hockly, 2015). In the context of Turkey, online exams became prevalent since 2015 in distance education centers. Ilgaz and Afacan Adanir (2019) performed an analysis of online exams comparing them with traditional exams in literacy, foreign language and history courses in addition to surveying perception of students about online exams. Their results showed that there were statistical differences between online midterm exams and traditional final exams of said courses. Hollister and Berenson (2009) conducted research to find differences between proctored and unproctored test performances of groups of students and found no difference regarding performance. Rios and Liu (2017) have also indicated the extensive use of unproctored exams in online education due to financial and flexibility issues while focusing on necessity of online proctoring facilities. In this sense, it might be important to balance available resources and desired outcomes. Recently, unproctored and proctored online exams have also gained popularity due to the COVID-19 pandemic.

2.1 Advantages and Disadvantages of Unproctored Online Exams (UOEs)

Online exams, specifically UOEs create new advantages that are not available in traditional exams. First of all, grading process can be automatically completed in online exams. In this sense, students are able to access their exam scores immediately, and receive feedback if given. Online exams can be reutilized with just a few clicks and eliminate all processes of exam preparation and printing that is carried out in traditional exams; therefore, it is indeed cost and time efficient. UOEs hold a crucial role in which students can access exam platforms with ease through low level requirements. Overall, UOEs not only provide advantages for learners but also for educators.

Although there are some major advantages in online exams of language courses, there are naturally some disadvantages. First of all, it may be difficult for instructors to set up question pools to be used in online exams as well as training instructors and learners for online exams (Clark et al, 2020). In transition from face-to-face traditional assessment to online assessment, academics need to review their assessment considerations and techniques (Hollister & Berenson, 2009). All this might be a time-consuming process for instructors and may require additional support. A contemporary server with adequate network bandwidth, CPUs and physical memory may be required. All these processes may increase initial costs of setting up an online exam environment. Additional disadvantages of these exams could be added stress on some learners who do not feel confident with technology. Goertler and Gacs (2018) similarly assert being successful in online language learning also depends on being competent in technology to some extent. This may affect overall assessment process in terms of validity.

Reliability and validity are the first issues to be considered in exams. Dermo (2009) indicates when an assessment is marked by computers, reliability of a test may increase which can be considered a further advantage of online assessment. Similarly, the reliability of a test results from producing statistically consistent measures when sampling error is eliminated (Dennick, Wilkinson, & Purcell, 2009). Therefore, the means and standard deviations of the grades of students registered at the distance education center subject to this study should be within acceptable boundaries affecting 20% of overall grade of students. UOEs with randomized questions require intense initial effort in order to maintain reliability and validity. Nonetheless, it may leave validity issues unanswered to some extent even though a qualified committee prepare question banks. In this sense, validity is a critical issue in all kinds of exams whether paper-based or online.

3. Method

3.1 Research Design

It is not always possible to manipulate independent variables in studies, yet natural conditions may take place such as a pandemic affecting dependent variables (Schenker & Rumrill, 2004). That is, causal-comparative research tries to "find relationships between independent and dependent variables" resulting from an action which is not possible to occur in normal conditions due to ethical considerations and regulations (Salkind, 2010, p. 124). In usual settings, UOEs are commonly used at distance education centers in Turkey; however, unproctored online final exams (UOFEs) were unexpected which were caused by the pandemic. Nonetheless, it may have less influence on distance education programs in comparison to traditional programs on campuses that have undergone a transition due to the COVID-19 pandemic. Since the following groups as categorical variables are easily available, a causal-comparative design is adopted in overall design of this research: different educational programs, gender, academic year, paper-based proctored final exams, unproctored online midterm exams (UOMEs), and UOFEs. The research questions that shape this study are presented below:

RQ1: Do students in online language courses have higher achievement UOMEs when compared to proctored paper-based finals exams?

RQ2: Do students in online language courses have higher achievement in UOMEs when compared to UOFEs?

RQ3: Do students in online language courses have higher achievement in proctored paper-based final exams of 2019 prior to the COVID-19 pandemic when compared to UOFEs of 2020 during early stages of the COVID-19 pandemic?

RQ4: Do students in online language courses have statistically different achievement scores in UOMEs when compared to proctored paper-based final exams across different educational programs?

RQ5: Do students in online language courses have statistically different achievement scores in UOMEs when compared to proctored paper-based final exams across gender?

3.2 Participants

The participants of the study were students of online programs at a public university in Turkey. The data collected through convenience sampling has been given below:

Table 1. *Demographic Information of Participants of 2019 and 2020 Spring Semesters*

	Variables	Frequency		Percentage (%)	
Year	2019-2020	2019	2020	2019	2020
Gender	Female	70	72	66.7	62.1
	Male	35	44	33.3	37.9
Educational Program	JSTC	8	15	7.6	12.9
	BAI	9	20	8.6	17.2
	CP	21	27	20.0	23.3
	LOMS	14	13	13.3	11.2
	MDS	46	37	43.8	31.9
	THBA	7	4	6.7	3.4
Total		105	116	100	100

Notes: JSTC: Justice; BAI: Banking and Insurance; CP: Computer Programming; LOMS: Law Office Management and Secretarial; MDS: Medical Documentation and Secretarial; THBA: Tourism and Hotel Business Administration.

3.3 Procedures

Twenty multiple-choice questions were selected randomly from question pools and appeared on the screens of learners in two online language courses at college level. In spring semester of 2019, final exams were taken in the form of paper booklets and exam results were obtained from optical answer sheets. Data was gathered from Moodle database in terms of exam score, gender and educational program. Once data was consolidated in spreadsheets for SPSS analyses, all details of student numbers and names were removed. Ethical procedures were followed by acquiring permission from the institutional review board of the Middle East Technical University, and the participants' rights and confidentiality were protected.

3.4 Data Analysis

Data were analyzed by using SPSS and utilizing nonparametric tests, as data did not adhere to normal distribution, and transforming data did not lead to normal distribution. Wilcoxon Signed-Ranks Tests, Mann–Whitney U Tests, and Kruskal-Wallis Tests were employed to answer corresponding research questions.

4. Results

The quantitative data from Moodle LMS exam logs were analyzed through descriptive statistics and presented by providing mean, standard deviation, and frequencies in the form of tables. Means and standard deviations of two courses of 2019 and 2020 spring semesters can be found in two tables below:

Table 2. *Comparison of Language Exams for Six Educational Programs of 2019*

Educational Program	Turkish Language Midterm Exam (Online & Unproctored)		Turkish Language Final Exam (Paper-based & Proctored)		English Language Midterm Exam (Online & Unproctored)		English Language Final Exam (Paper-based & Proctored)	
	M	SD	M	SD	M	SD	M	SD
JSTC	76.87	7.98	81.25	9.54	82.50	28.41	69.37	26.38
BAI	66.66	13.69	66.11	14.52	63.33	16.39	27.77	21.37
CP	72.85	21.71	76.66	13.07	83.80	14.73	66.90	21.18
LOMS	70.35	19.46	65.71	11.57	61.42	18.75	29.28	15.42
MDS	72.93	14.51	71.08	11.10	72.71	19.93	47.82	18.96
THBA	54.28	13.97	53.57	15.19	67.14	24.64	42.85	20.17

Table 3. *Comparison of Language Exams for Six Educational Programs of 2020*

Educational Program	Turkish Language Midterm Exam (Online & Unproctored)		Turkish Language Final Exam (Online & Unproctored)		English Language Midterm Exam (Online & Unproctored)		English Language Final Exam (Online & Unproctored)	
	M	SD	M	SD	M	SD	M	SD
JSTC	78.00	10.48	74.00	9.67	73.33	19.88	73.00	22.02
BAI	63.25	15.83	63.25	12.69	69.50	29.15	67.50	24.52
CP	77.96	12.34	77.22	12.03	82.96	17.05	84.81	15.09
LOMS	62.30	16.28	68.46	11.43	53.84	22.92	57.69	25.46
MDS	70.54	12.29	67.83	13.92	70.13	18.04	69.18	21.13
THBA	62.50	17.07	57.50	13.22	88.75	14.36	70.00	19.57

4.1 Achievement Across Unproctored Online Midterm (UOME) and Proctored Final Exams

To evaluate statistical difference between UOME and proctored paper-based final exam scores of Turkish language course in the spring semester of 2019 a Wilcoxon Signed-Ranks Test was carried out which indicated that there were no statistically significant differences between UOME scores and paper-based proctored final exam scores T = 2031, z = -.247, p = .805. Similarly, for the English language course in the spring semester of 2019 a Wilcoxon Signed-Ranks Test indicated that there were statistically significant differences between UOME scores and paper-based proctored final exam scores T = 300, z = -7.485, p < .001. That is, students performed worse in proctored paper-based final exam compared to UOEs of English language course in 2019 before COVID-19 pandemic.

4.2 Achievement Across Unproctored Online Midterm (UOME) and Unproctored Online Final Exams (UOFE)

To evaluate any statistical difference between UOME and UOFE scores of Turkish language course in the spring semester of 2020 a Wilcoxon Signed-Ranks Test was run and showed that there was no statistically significant difference between UOME and UOFE scores T = 2444, z = -.446, p = .656. Similarly for the English language course in the spring semester of 2020 the same test showed there were no statistically significant differences between UOE midterm scores and UOE final scores T = 2437, z = -.134, p = .894. That is, students performed similarly in both cases of UOE instances at different times during the pandemic.

4.3 Achievement Comparison Prior and During the COVID-19 Pandemic

For exam scores for the years 2019 and 2020, Mann–Whitney U tests indicated that the achievement scores only statistically differed between proctored paper-based English language and UOFE , $U(N_{2019} = 105, N_{2020} = 116) = 44.18$, $p < .001$. The median of the English language final exam achievement scores in 2020 (*Md* = 80) was higher than the median in 2019 (*Md* = 50).

4.4 Achievement Across Different Educational Programs

For the UOME for the year 2019, a Kruskal-Wallis test revealed that there was no statistically significant difference in the UOME scores of the Turkish language course across six educational programs, $H(5) = 10.57$, $p = .061$. Medians of exam scores across different educational programs were, to some extent, similar and were JSTC (*Md* = 80), BAI (*Md* = 70), CP (*Md* = 80), LOMS (*Md* = 68), MDS (*Md* = 75), THBA (*Md* = 55). However, when it was the proctored paper-based final exam for the year 2019, a Kruskal-Wallis test revealed a statistically significant difference in proctored paper-based final exam scores of the Turkish language course across six educational programs, $H(5) = 21.55$, $p = .001$. Two pairs of educational programs according to Dunn's pairwise tests statistically differed. JSTC-THBA and CP-THBA were these pairs, and there was strong evidence ($p = .005$, adjusted using the Bonferroni correction) between JSTC and THBA, and ($p = .002$) between CP and THBA. JSTC (*Md* = 82) and CP (*Md* = 80) performed better than THBA (*Md* = 60).

For the UOME for the year 2020, a Kruskal-Wallis test revealed a statistically significant difference in UOME scores of the Turkish language course across six educational programs, $H(5) = 21.16$, $p = .001$. Dunn's pairwise tests indicated two pairs of educational programs statistically differed which are CP-LOMS and CP-BAI. There was some evidence ($p < .05$) for both between CP-LOMS and CP-BAI, and CP (*Md* = 85) performed better than BAI (*Md* = 65) and LOMS (*Md* = 60). Similarly, for the UOFE for the year 2020, the same test revealed a significant difference in UOFE scores of the Turkish language course across six educational programs, $H(5) = 18.55$, $p = .002$. Only one pair of educational programs statistically differed, and it took place between CP and BAI. There was somewhat strong evidence ($p < .01$) between CP and BAI, and CP (*Md* = 80) performed better than BAI (*Md* = 62).

For the UOME for the year 2019, a Kruskal-Wallis test revealed a statistically significant difference in UOME scores of the English language course across six educational programs, $H(5) = 16.53$, $p = .005$. Only one pair of educational programs statistically differed, CP and LOMS. There was some evidence ($p = .020$) between CP and LOMS, and CP (*Md* = 85) performed better than LOMS (*Md* = 65). Similarly, for the proctored paper-based final exam for the year 2019, the same test indicated a significant difference in proctored paper-based final exam scores of the English language course across six educational programs, $H(5) = 34.15$, $p < .001$. Four pairs of educational programs differed, CP-BAI, JSTC-BAI, CP-LOMS and JSTC-LOMS. There was strong evidence ($p < .005$) between these pairs, and CP (*Md* = 70) and JSTC (*Md* = 78) performed better than BAI (*Md* = 20) and LOMS (*Md* = 25).

For the UOME for the year 2020, a Kruskal-Wallis test revealed a statistically significant difference in UOME scores of the English language course across six educational programs, $H(5) = 18.01$, $p = .003$. Only one pair of educational programs differed, CP and LOMS. There was strong evidence ($p < .005$) between CP and LOMS, and CP (*Md* = 80) performed better than LOMS (*Md* = 70). Similarly, for the UOFE for the year 2020, the same test indicated a statistically significant difference in proctored paper-based final exam scores of the English language course across six educational programs, $H(5) = 16.25$, $p = .006$. Two pairs of educational programs statistically differed which are CP-LOMS and CP-MDS. There was some evidence ($p < .05$) for both between CP-LOMS and CP-MDS, and CP (*Md* = 90) performed better than LOMS (*Md* = 45) and MDS (*Md* = 70).

4.5 Achievement Across Gender

Mann–Whitney U tests for gender influences on each of all four exams of English language courses showed no statistically significant differences except in one of the UOFEs in 2020. At this time, the Mann-Whitney U test revealed that UOFE scores of male students (*Md* = 88, n =44) in English language course were higher than ones of female students (*Md* = 70, n =72), $U = 1118$, $z = -2.66$, $p = .008$.

5. Discussion and Conclusion

The literature review shows that online language learning and online assessment, specifically UOEs, provide challenges and opportunities for both learners, instructors and institutions. This poses many comparably new phenomena to consider especially for stakeholders who do not have experience in online learning and teaching. Creating reliable and valid question categories, preventing cheating, and easing examination process for all are just some of concerns. As this study focused on UOEs in online language courses, it revealed some varying results depending on different research questions. In only one case, students had higher achievement in UOMEs when compared with proctored final exams, and it took place between UOME and proctored paper-based final exam of online English language course in 2019. Although this could be regarded as a result of cheating, it cannot be certainly known and asserted due to the limitations of this study. When both midterm and final exams were held in uproctored and online environments, there were no statistical differences between midterm and final exams of each language course. When two academic semesters prior and during COVID-19 pandemic were compared, it was understood that students had higher achievement scores in unproctored online English language final exam during the pandemic. Educational programs of CP and JSTC continuously showed higher achievement results in different situations. Gender had an impact on achievement of unproctored English language exams during early period of COVID-19 and the situation was in favor of male students. In the case of UOEs, it may be further necessary to detect gender influences in different domains in order to address underlying issues. That is, online learning environments and online learner populations continue to increase, research studies are to be conducted in different domains and by taking into account different considerations. Further research interests related to online exams may focus on carrying out investigations on enabling safe exam browsers, disabling navigation of questions, disabling text-copy feature, and utilizing proctoring software during exams such as screen, voice, and camera recording, all of which were not utilized in UOEs of this study. Regardless, there appears to be many advantages of UOEs indicated across the literature. Therefore, it is highly likely that UOEs will be utilized in the future, especially for online midterm exams in distance education programs with cost-efficient examination designs in which overall achievement of a student taking a course does not solely depend on passing an online exam.

References

Clark, T. M., Callam, C. S., Paul, N. M., Stoltzfus, M. W., & Turner, D. (2020). Testing in the Time of COVID-19: A Sudden Transition to Unproctored Online Exams. *Journal of Chemical Education*, *97*(9), 3413–3417.

Dennick, R., Wilkinson, S., & Purcell, N. (2009). Online eAssessment: AMEE Guide No. 39. *Medical Teacher*, *31*(3), 192–206.

Dermo, J. (2009). E-Assessment and the student learning experience: A survey of student perceptions of e-assessment. *British Journal of Educational Technology, 40*(2), 203-214.

Goertler, S., & Gacs, A. (2018). Assessment in online German: Assessment methods and results. *Die Unterrichtspraxis/Teaching German, 51*(2), 156-174.

Hockly, N. (2015). Developments in online language learning. *ELT Journal, 69*(3), 308-313.

Hollister, K. K., & Berenson, M. L. (2009). Proctored versus unproctored online exams: Studying the impact of exam environment on student performance. *Decision Sciences Journal of Innovative Education, 7*(1), 271-294.

Ilgaz, H., & Afacan Adanır, G. (2019). Providing online exams for online learners: Does it really matter for them? *Education and Information Technologies, 25*(2), 1255-1269.

Rios, J. A., & Liu, O. L. (2017). Online proctored versus unproctored low-stakes internet test administration: Is there differential test-taking behavior and performance? *American Journal of Distance Education,* 1-14.

Salkind, N. J. (2010). *Encyclopedia of research design*. Los Angeles, CA: SAGE.

Schenker, J. D., & Rumrill, P. D. (2004). Causal-comparative research designs. *Journal of Vocational Rehabilitation, 21*(3), 117-121.

Supporting System to Encourage Self-review in Composition Class — Investigation into the Learner's Reaction

Yan ZHAO*, Haruhiko TAKASE & Hidehiko KITA
Graduate School of Engineering, Mie University, Japan
**419DE52@m.mie-u.ac.jp*

Abstract: Composition class to learn a foreign language is effective but has many issues. In this study, we aimed to encourage learners to naturally review their own writing by using a system that points out where they have made errors. The support system provides learners with feedback on where they have made errors. We investigated revising status in composition class using the system. As a result, we clarified that learners who use the system can actually correct their own errors though the system's feedbacks are not completely accurate.

Keywords: Second language learning, composition class, error checker

1. Introduction

In second language learning, there are four skills to be acquired: listening, speaking, reading, and writing. In this study, we focus on writing skill. To acquire writing skills, there is various types of classes. We focus on composition class, especially Japanese learning as second language. Teachers should correct many compositions written by learners in composition classes. It is burden on the teacher.

Many systems to support composition classes have been developed. Some studies have aimed to encourage learners' writing activities. Some of them try to pointing out errors in writing. Sato (2014) stated that many students do not want to be given the correct answer directly by the teacher, but want to be given feedback, using symbols, and so on. Therefore, it is important that feedbacks are given as indirect feedback so that students can notice errors. Therefore, Zhao (2021) focused on a system that encourages learners to revise based on the information of detected errors.

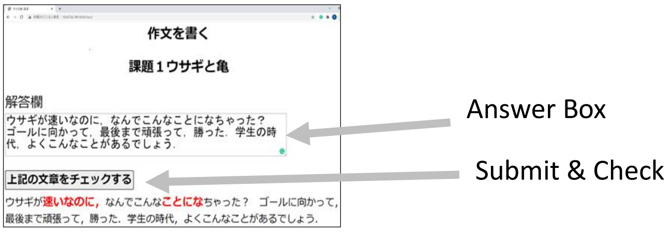
In this study, we investigate the learner's reaction to the feedback of Zhao's system (see Zhao (2021)). Since the system feedbacks provide limited information (only the location of errors) and some feedbacks are not accurate, we investigate whether learners correct their own composition with the system's support or not.

2. System that Encourages Self-review

We simply explain Zhao's system (see Zhao (2021)). Figure 1 shows an overview of the screens of the system. The screen is divided into the following three major parts.

1. Textbox to write a composition
   The composition should be typed in the space below the "Answer Box".
2. Submit button
   By pushing the button "Submit & Check", the system starts to check the inputted composition.
3. Area for the system's feedbacks
   Below the submit button, the system's feedbacks are displayed. They are showing the location of errors in large bold red letters.

First, a learner inputs the composition into the textbox, and pushes the submit button. Next, the system checks it in the background. The checker is based on Zhao's method (see Zhao (2020)), which detects grammatical errors that can be judged from only one clause. Then, the learner receives the results, which are only the location of errors, and revises the composition. The learner repeats this process and submits the composition to the teacher.



作文を書く

課題１ウサギと亀

解答欄

ウサギが速いなのに，なんでこんなことになちゃった？ゴールに向かって，最後まで頑張って，勝った．学生の時代，よくこんなことがあるでしょう．

ウサギが**速いなのに，**なんでこんな**ことにな**ちゃった？ ゴールに向かって，最後まで頑張って，勝った．学生の時代，よくこんなことがあるでしょう．

Figure 1. Screen for Learners

3. Problems

The system shows the locations of errors in conspicuous letter. We should pay attention that the feedback from the system is not completely accurate. Because the system checks compositions with the check rules that were automatically acquired from limited data by using machine learning techniques. Some learners may not rely on all system's feedback because of only a few inaccurate feedbacks. Some of the others may completely rely on all feedbacks and fail to correct all errors.

4. Experiment

In this section, we discuss that learners correct their own composition or not according to inaccurate feedbacks by simple experiments. There are two points to be checked. (1) Can learners revise their compositions only from the system's feedback? (2) Can learners find errors that the system missed?

4.1 Experimental Conditions and Procedures

We conducted an experiment that was a composition class using the system with seven Chinese learners of Japanese, whose level of the International Japanese Language Proficiency Test (ILPT) of learners are as follows: 3 N1 level learners, 2 N2 level learners, and 2 N3 level learners. They are 20-40 years old and had been living in Japan for more than 2 years. Each learner wrote a composition, which was 200-300 words, based on the content of a four-panel manga (The Hare and the Tortoise). The procedure for this composition class is shown below.

Step 1: Before starting the composition, each learner receives guidance for using the system, and answers a pre-assessment questionnaire (20 minutes).
Step 2: The learner writes a composition by looking at the four-panel manga, receives feedbacks for the composition from the system, and revises the composition (55 minutes).
Step 3: The learner revises he/she compositions according to the teacher's suggestions.
Step 4: The learner answers a post-questionnaire (10 minutes)

4.2 Results and Discussion of the Experiment

First, we discuss the first issue: "Can learners revise their compositions only from the system's feedback?". The system pointed out the 213 errors for all compositions. Learners corrected 59 errors of 213 errors and revised 43 errors of 59 errors correctly. It means that only 28% of errors, which were pointed out by the system, were fixed by learners. But the system's feedback was not completely accurate: only 77 errors of 213 errors were accurate. So, learners select suitable feedback and revise them (43 errors of 77 errors = 56%) without the teacher's suggestion. Consequently, learners can revise their compositions only from the system's feedback.

Next, we discuss the second issue: "Can learners find errors that the system missed?". The system missed 34 errors to point out. But learners found 3 errors of 34 errors and revised them correctly. It means that learners inferred similar errors and found some errors by themselves. It does not mean that learners can find errors without the systems. Learners find errors, since the system stimulated the learners' cognitive activities.

Table 1. *Detail of Review*

	Learners							
	A	B	C	D	E	F	G	Total
Over all corrected what was pointed out to them	67	56	57	57	83	50	41	56
Number of times the learner corrected something that was not pointed out	67	20	0	0	0	0	0	9
Corrected the correct place pointed out and the place pointed out incorrectly	0	0	0	0	0	0	0	0
								(%)

For additional discussion, Table 1 shows the summary of each learner's revision status of the compositions after step 2. On average, 56% error of the system's suitable feedbacks was fixed by learners. In the viewpoint of each learner, the ratio is from 41% to 83%. Since the worst learner fixed 41% of errors, the system's feedback would be useful for all most learners. Only 9% errors of uncommented by the system were found by only some learners. Since it would depend on the learner's language skill, it is the natural result. Finally, all 136 errors, which were miss detected errors by the system (213-77 errors), were not fixed by learners. Though the system cannot check compositions completely accurately, learners can judge suitable feedback and fix errors.

Consequently, learners revise their compositions in an inaccurate system. In other words, learners can select correct system feedbacks by themselves and find similar errors which are not pointed out.

5. Conclusion

In this study, we discussed the effectiveness of the system that points out the location of errors to learners of Japanese. We conducted a simple experiment to confirm that learners can revise their errors in composition with the system's feedback. We showed the following results: (1) learners revise 56% errors of system's suitable feedbacks. It means that learners can revise their compositions only from the system's feedback. (2) learners revise 9% errors of the system missed. It means that learners inferred similar errors and found some errors by themselves. These results suggest that the system is not complete but useful for learners in the composition class.

In the future, we will discuss how to reduce the number of incorrect system's feedbacks and conduct further experiments to show the effectiveness of the system.

Acknowledgements

Part of this work was supported by the Mie University Engineering Alumni Association Research Grant.

References

Sato, S. (2014). A study of feedback in English output activities: from the perspective of intrinsic motivation. *Bulletin of the Shikoku Society for English Language Education*, (34), 67-78.

Zhao, Y., Takase H., & Kita H. (2020). Detection of errors in Japanese learner's composition by machine learning —Detection of grammatical errors in a clause—. *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, 32(5), 887-890 (in Japanese).

Zhao, Y., Takase H., & Kita H. (2021). Support system to encourage student's self-review in composition class for second language learners. *Computer & Education*, 50, 96-99 (in Japanese).

Examining the Effects of Automatic Speech Recognition Technology on Learners' Lexical Diversity

Michael Yi-Chao JIANG*, Morris Siu-Yung JONG & Wilfred Wing-Fat LAU
Department of Curriculum and Instruction & Centre of Learning Sciences and Technologies, The Chinese University of Hong Kong, Hong Kong S. A. R., China
*mjiang@cuhk.edu.hk

Abstract: A total of 160 undergraduates participated in the 14-week quasi-experiment. The experimental group and the control group were both taught with a flipped approach, but the students in the experimental group were required to conduct an additional automatic speech recognition-based pre-class task. The vocd-D value and MTLD were adopted as metrics of students' lexical diversity. A two-way between- and within-subjects repeated measures design was conducted to examine the effects of the group factor, the time factor and the group × time interaction effects. The results showed that the students in the experimental group scored statistically better than their counterparts in the control group on both the vocd-D value and MTLD. However, no significant difference was witnessed over time and there was no significant group × time interaction effect in either group.

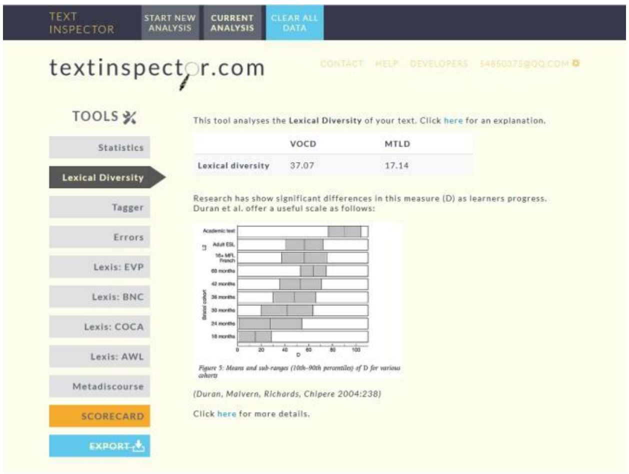
Keywords: Automatic speech recognition, flipped classroom, lexical diversity, English as a foreign language

1. Introduction

In flipped learning, students are supposed to prepare themselves with the pre-class content actively and achieve a proper level of preparedness. However, it was revealed that educational technology was not fully harnessed in flipped classrooms (Jiang et al., 2020). Advanced technologies such as automatic speech recognition (ASR) are not commonly considered and utilized. Because of the limited use of technology in some flipped classrooms, students may become as demotivated as do they in traditional classrooms. In EFL research, lexical diversity is considered as a basic descriptor of learners' oral English proficiency. Although many extant studies have reported positive evidence of the effectiveness of the FCA in EFL learning, few studies have adopted domain-specific indicators (e.g., specific measures of students' oral fluency or accuracy) to examine the effectiveness of the technology-enhanced flipped classroom approach (FCA) (Jiang et al., 2021). Therefore, the present study aims to examine the effects of ASR technology on EFL learners' lexical diversity in a flipped setting. Studies with more refined indicators of learners' linguistic performance can contribute to a deeper understanding of the FCA for language learning and diversify the instructional design of the FCA. Accordingly, we formulated two research questions: 1) Do students in a flipped EFL classroom outperform their counterparts in a traditional EFL classrooms in terms of lexical diversity? 2) How does the flipped classroom approach improve the EFL students' lexical diversity over time?

2. Methods

A total of 160 first-year students from a four-year university were enrolled in the main study. Before the course began, the four classes had been randomly assigned into an experimental group (EG) and a control group (CG). The EG students were assigned a mediating ASR-based oral task in addition to the self-learning resources on Unipus for pre-class preparation. In contrast, the CG students were only given the materials on Unipus before class. All the students were randomly assigned to workgroups of two to four for in-class activities within each class. For each unit, the students in both groups undertook a communicative task in class in English and recorded the whole activity using their mobile phone in an auditory fashion. Within each workgroup, the students orally expressed their opinions or experiences regarding the unit topic. The recordings of Units 2, 4, 6 and 8 were used for data analysis, but the students did not know which unit would be analyzed. After data screening, the present study conducted data analyses based on the data that were transcribed and coded from a total of 128 participants (68 EG students and 60 CG students). The in-class recordings were transcribed into plain text and coded and annotated using ELAN (https://tla.mpi.nl/tools/tla-tools/elan). The vocd-D value and MTLD were computed with TextInspector (Figure 1), a professional website for studying linguistic features of English spoken language. The transcribed recordings of in-class peer interaction were coded into frequencies and relative frequencies (against AS-unit) to form study-generated quantitative data. A mixed within- and between-subjects design (pre-intervention English proficiency controlled for as a covariate) was conducted. The independent variables were the group factor and the time factor. The dependent variables were the metrics of lexical diversity coded from students' in-class task performance.



This tool analyses the **Lexical Diversity** of your text. Click here for an explanation.

	VOCD	MTLD
Lexical diversity	37.07	17.14

Research has show significant differences in this measure (D) as learners progress. Duran et al. offer a useful scale as follows:

Figure 5: Means and sub-ranges (10th–90th percentiles) of D for various cohorts

(Duran, Malvern, Richards, Chipere 2004:238)

Click here for more details.

Figure 1. Screenshot of TextInspector for Calculating vocd-D and MTLD.

3. Results and Discussion

The means of vocd-D value over time was 41.716 for the EG students and 36.483 for the CG students. The means of MTLD over time was 32.807 for the EG students and 28.060 for the CG students. In terms of the timepoints, the means of the vocd-D values across the two groups was 35.552 for Time 1, 36.009 for Time 2, 39.782 for Time 3, and 44.401 for Time 4, showing a clear upward tendency (Figure 2). The means of MTLD across the two groups was 28.700 for Time 1, 29.463 for Time 2, 28.497 for Time 3 and 31.958 for Time 4, indicating an upward yet fluctuating pattern. The test of between-subjects effects revealed that the main effect of the group factor on the average score of vocd-D values across time was significantly different ($F_{(1,\ 125)}$ = 3.945, p = 0.049 < 0.05) and the effect size was small-to-medium (η^2_{p} = 0.031 > 0.01). Likewise, the main effect of the group factor on the average score of MTLD across time was also significantly different ($F_{(1,\ 125)}$ = 6.244, p = 0.014 < 0.05) also with a small to medium effect size (η^2_{p} = 0.048 > 0.01). Therefore, the EG students performed significantly better than their counterparts in the CG in terms of lexical diversity.

The tests of within-subjects effects showed that the main effect of time was not statistically significant on the average scores on vocd-D values ($F_{(2.475,\ 309.351)}$ = 0.483, p = 0.658 > 0.05) or on MTLD ($F_{(2.481,\ 302.224)}$ = 0.049, p = 0.971 > 0.05), sphericity not assumed. Additionally, the 'group × time' interaction effects on vocd-D value ($F_{(2.475,\ 309.351)}$ = 2.720, p = 0.055 > 0.05) or on MTLD ($F_{(2.481,\ 302.224)}$ = 0.827, p = 0.458 > 0.05) were not statistically significant, sphericity not assumed. To sum up, the time factor did not lead to any statistically significant effects on lexical diversity, and there was no group × time interaction effect on neither of them.

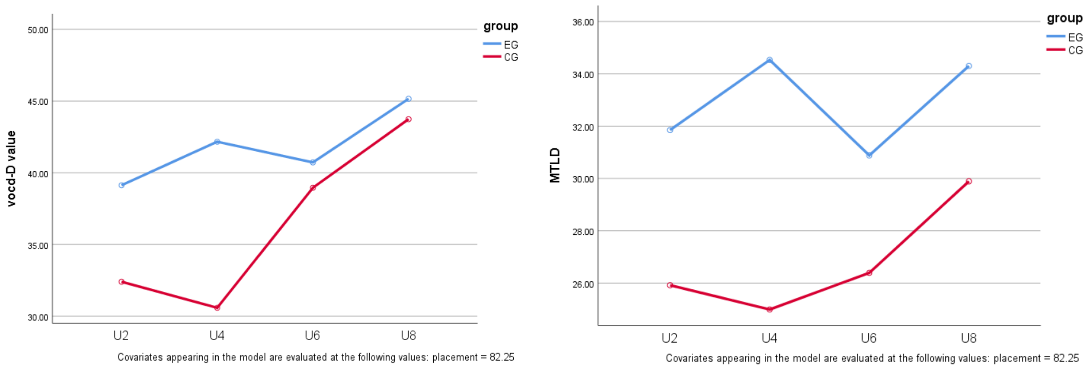


Figure 2. Profile Plots of vocd-D Value and MTLD.

The significant gains on lexical diversity indicated that ASR technology had positive effects on EG students' in-class linguistic performance in terms of lexical diversity in oral English. The EG students were required to use the ASR technology-based iFlyRec to practice their English speaking skills for their pre-class self-learning. They were encouraged to repeatedly practice answering the lead-in questions attached to the reading sections. The students needed to master the new vocabulary and have a good knowledge of the text contents to answer the lead-in questions properly, which to some extent, reinforced their mastery of and facilitated the use of the new words. In contrast, their counterparts in the CG did not have any required practice of using the new vocabulary. According to the responses in the post-intervention interview, it was found that most of the students self-studied the new vocabulary mostly by memorizing by rote, focused only on building meaning connections between Chinese and English, spelling and pronunciation. They had little awareness in learning the usage of the new words. Therefore, when performing the communicative tasks in class, the EG students scored statistically more on the lexical diversity dimension. Students' lexical diversity was not enhanced significantly over time. These resultant contrasting findings align with most of the complexity, accuracy and fluency (CAF) studies conducted among non-native speakers (e.g., Skehan, 2009). It was found in those studies that the correlation between lexical diversity and syntactic complexity was shown to be negative, indicating that for non-native speakers, 'more varied lexis seems to cause problems for non-native speakers and provokes more errors while not driving forward lexical diversity' (Skehan, 2009, p. 116).

4. Conclusion

The results revealed that the EG students scored statistically higher on both vocd-D and MTLD than the CG students, indicating that the EG students produced more complex utterances on the lexical level than their counterparts in the CG. However, regardless of their group membership, students' lexical diversity did not improve significantly over time. In other words, the ASR technology significantly improved the EG students' lexical diversity, but over time, the lexical diversity of both the EG students and the CG students did not improve significantly.

References

Jiang, M. Y. C., Jong, M. S. Y., Lau, W. W. F., Chai, C. S., Liu, K. S. X., & Park, M. (2020). A scoping review on flipped classroom approach in language education: Challenges, implications and an interaction model. *Computer Assisted Language Learning.*

Jiang, M. Y. C., Jong, M. S. Y., Lau, W. W. F., Chai, C. S., & Wu, N. (2021). Using automatic speech recognition technology to enhance EFL learners' oral language complexity in a flipped classroom. *Australasian Journal of Educational Technology, 37*(2), 110-131.

Skehan, P. (2009). Lexical performance by native and non-native speakers on language learning tasks. In B. Richards, H.M. Daller, D. Malvern, P. Meara, J. Milton, & J. Treffers-Daller (Eds.), *Vocabulary studies in first and second language acquisition: The interface between theory and application.* UK: Palgrave Macmillan.

Technology Integration in a Communicative English Classroom

Ruth Z HAUZEL
GITAM (Deemed to be University), Hyderabad, India
rhauzel@gitam.edu

Abstract: The purpose of this study is to analyse the integration of technology and its impact using Moodle activities from a teacher's perspective. This paper is a case study involving 59 Bachelor of Technology, 1st-year second semester students in a Communicative English course who were taught the same course in a face-to-face mode before the pandemic and then switched to online classes. Although several studies have focused on analysing the use of the Moodle platform as a whole, few have examined the use of each activity included and its potential impact on learning. Analysis and findings from the case study suggest that assignments, quizzes, and lessons are the activities that have a major learning impact and provides new educational scenarios. On the other hand, forum activity (communication tool), could not be utilised to its potential.

Keywords: Moodle activities, online learning, teacher perspective, teacher training, technology integration

1. Introduction

One of the most commonly used Learning Management Systems (LMS) for developing online academic courses is Moodle. The main feature that differentiates the traditional learning environment and using LMS such as Moodle is the degree of technology usage and the gradual shift of control and responsibility of the learning process to the learners. Moodle gives learners the opportunity to learn anytime, anywhere and it favours a student-centered approach with teachers acting as 'organizers, advisers, and sources of information' (Horváth 2007: 104).

2. Context of the Study

This disruption necessitated by the pandemic is unprecedented as teaching and examination has delayed student progression and this resulted in the sudden shift from traditional classroom teaching to online teaching. Keeping this context in mind, the study was conducted for one semester, November 2019-April 2020 in a Communicative English classroom. Weekly there were two Communicative English lab sessions for two hours each and two classroom teaching hours for fifty minutes each, a total of six hours of Communicative English class in a week. Different tasks such as reading comprehension, listening skills, grammar and vocabulary were taught through the platform.

3. Literature Review

A significant feature of using Moodle is that it optimizes the teaching-learning process besides being an effective and flexible learning environment for learners. A study conducted by Escobar-Rodriguez and Mongo- Lozano (2012) showed that the learning–teaching process is improved, and students obtain better skills and grades by using Moodle. It also suggests that Moodle makes the whole learning process more interesting and friendlier (Martín-Blas and Serrano-Fernández, 2009). It was also found that most information technology majors perceive learning to be more fun and of better quality within a technology-enhanced online learning environment (Parker, 2003). However, lack of interaction, presence, or both may result in students' different observations on how well they may or may not have performed in an online class (Song, Singleton, Hill, & Koh, 2004). More recent studies include new learning benefits related to collaborative work, learning outcomes, learning interest and creativity, and learning strategies for the students (Petko 2012).

The implementation of constructivist notions of theory into practice has been attempted in many learning environments, and most recently in technology and higher education (Doolittle, 1999; Roth & Lee, 2007). Vygotsky's cultural-historical theory of psychological development informed the foundation of sociocultural theory and constructivist practices of teaching and learning (Kozulin, 1998; Vygotsky, 1978; Wells, 1999; Wenger, 1998). The technological design of Moodle underpins the social constructionism theory where learning is considered a process of constructing knowledge by negotiating meaning with others and creating shared cultural artifacts. It is a learning-centred management system which draws on the *social constructivist framework* (Duffy & Cunningham 1996; Williams & Burden 1997). When integrating technology as a tool for learning, the following are assumptions for designing contemporary pedagogical practices infused with constructivist theory in classrooms that view: (a) learning as a process of construction so there will be multiple constructions/perspectives, (b) learning in contexts that are relevant to the learner, (c) learning mediated by tools (technology) and signs (semiotic tools), and (d) learning as a social-dialogical activity (Duffy & Cunningham, 1996; Vygotsky, 1978). An LMS is needed to support constructivist theory with pedagogical recommendations. Moodle can provide a unique opportunity for students to engage in social negotiation and mediation in the form of asynchronous (e-mail, threaded discussions) and synchronous (simulations, web-based data collection, and ill-structured problem solving) technology.

4. Need for the Study

The sudden shift from conventional pedagogy to online learning due to the pandemic, requires teachers to adapt and integrate technology for online teaching overnight with little awareness or training. Students were already introduced to the Moodle platform before the lockdown but it was used only as a supplement to face-to-face classroom teaching. However, Moodle suddenly became the only platform for all activities, assignments, quizzes, sharing notes, etc once lockdown was imposed.

5. Participants

The study is based on a data collected from 59 1st year second semester students who are pursuing their Bachelor of Technology course. These 59 Computer Science students are from the same section and are part of a compulsory Communicative English course. All the students are from English medium school who have had an exposure to English language for more than 10 years. The class consist of 23% girls and 76.8% boys, and all the students are between the age group of 17-19 years.

6. Data Analysis

From the online survey administered it was found that 75% of the students access the internet many times a day, while 10.7% access several times a week, 8.9% access once a day, and only 5.4% once a week. It was found that in spite of the students using the LMS for the first time, 83.9% stated they are comfortable while 16.1% stated that they are not very good with it. Another interesting observation was the students did not have much difficulty in accessing course materials uploaded and 67.9% says they can easily navigate through the platform while, 30.4% finds it a little challenging. Though students were using LMS for the first time 80.4% enjoys learning using the platform and 17.9% says they are neutral and just okay and 1.8% says they did not enjoy learning using the platform. In line with what the study intends to analyse, 76.8% stated that the course activities helps them learn and also prepare for examinations better, 21.4% stated that it helps them to a certain extend only and 1.8% does not find it helpful.

The Moodle activities used in this study are quizzes, assignments, lessons, and forums. All these activities used either in the classroom or outside the classrooms were monitored using progress completion chart. This is very useful for teachers as well as students as it gives an indication of the activities that the students completes or missed. This also provides a holistic view on the topics completed and one need not make note of activity completion elsewhere. When it comes to Forum (discussion tool) activities, in spite of many advantages it was very challenging to monitor what students respond to each other, especially when the strength of the class is more than 50. The purpose of a discussion forum was to engage in a discussion just like the physical classroom discussion. However, students can get personal with their comments which could lead to unnecessary arguments or unpleasant situation.

After observing how students respond to each other in the discussion forum, there was a need to change how forum is used. After the total lockdown with only online teaching possible, forum discussion activity was altered and very specific tasks were given, even a discussion topic had specific pointers to be included. For example, students were given a reading passage and they have to post their response on the forum. A deadline for all such activities were given and students could complete the task at their own convenience within the stipulated time. It was observed that these kind of tasks were very useful in completing the syllabus and engaging students outside the classroom as well.

Another useful activity is the lesson activity which allows teachers to create branching exercises and upload content including multimedia. This activity can also be used to upload reading texts from a prescribed textbooks or any other sources. Quiz is another powerful tool that meets teaching needs as it can be used to check understanding of simple concept that has been recently taught or it can also be used as formative assessment. Students enjoyed taking quizzes as they were able to see their performance immediately and also get feedback for each question. Another benefit of this activity is that it records the submission details which eases the teacher’s burden of having to keep track of late submission of work.

7. Discussion and Finding

The non-productive engagement of students on the communication and collaboration tools such as forum, as well as the need to adapt the use of these activities, brings into question whether teaching practice using Moodle is really in line with the social constructionism theory of learning which underpins the technological design of Moodle. This finding may be attributed to students lack of interest or a negative perception of collaborative work unlike in a face-to-face mode. Also, findings may be due to teachers’ beliefs that students will most likely have a negative perception of collaboration using Moodle (Psycharis et al. 2013) in face-to-face scenarios. This belief makes it necessary to rethink the role that Moodle should realistically have in face-to-face educational setting, particularly since it was designed to develop communication in online scenarios.

After using different Moodle activities as shown above during a course of time, from a student’s perspective this shift of control of the learning process to the learners seems to positively influence their learning effectiveness. The results showed that students engagement level is also high. Overall, students had a positive attitude while working on the task and using the medium. However, from the above analysis and from a teacher’s perspective, the integration of technology is yet to have the desired impact in the classrooms. The successful use of such platforms in the teaching and learning context critically depends on the teachers having knowledge about the tools, being aware of how they should be used and being capable of organizing all the communication process. This calls for the need to systematically integrate technology over a period of time and that teachers need to be trained.

Another significant point to keep in mind is that studies indicated that the technology integration practices of teachers in the classroom often did not match their teaching styles. This could be due to external barriers that prevented teachers from using technology in ways that matched their practiced teaching style or it could be the lack of professional support, lack of in-service training, lack of available technology, restricted curriculum, training and above all time to practice how to use such tools. Another factor to consider is that increasing the amount of technology in the classroom was not sufficient to change teachers' technology practices without a shift in the teachers' pedagogical practices. Therefore, all these indicates the need to restructure the professional development on strategies for contextualizing technology integration in the classroom.

8. Conclusion

The results of this study suggest the need for teachers to exploit technological tools for professional development and the urgent need for training. While integration of technology is important in the education system, just providing access to technology is not adequate. Meaningful development of technology-based knowledge is significant for all learners in order to maximize their learning. The technology of today shortly becomes the technology of yesterday in education. Technology and its uses are constantly changing to incorporate new ways of managing interactions on the digital plain (Dede, 2010). Many teachers are still struggling to achieve meaningful technology integration within their classrooms and there are implications for practice, specifically related to the continued professional development because of the current situation.

References

Badia, A., Martín, D., & Gómez, M. (2019). Teachers' perceptions of the use of Moodle activities and their learning impact in secondary education. *Technology, Knowledge and Learning*, *24*(3), 483-499.

Bri, D., Coll, H., Garcia, M., Lloret, J., Mauri, J., Zaharim, A., ... & Kalogiannakis, M. (2008, July). Analysis and comparative of virtual learning environments. In WSEAS international conference. Proceedings. Mathematics and computers in science and engineering (No. 5). WSEAS.

Ertmer, P. (2005). Teacher pedagogical beliefs: The final frontier in our quest for technology integration? *Educational Technology Research and Development*, *53*(4), 25-39.

Ertmer, P. A., Ottenbreit-Leftwich, A. T., Sadik, O., Sendurur, E., & Sendurur, P. (2012). Teacher beliefs and technology integration practices: A critical relationship. *Computers & education*, *59*(2), 423-435.

Finnegan, M., & Ginty, C. (2019). Moodle and social constructivism: Is Moodle being used as constructed? A case study analysis of Moodle use in teaching and learning in an Irish higher educational institute. *All Ireland Journal of Higher Education*, *11*(1).

Kozulin, A. (1998). *Psychological tools: A sociocultural approach to education*. Harvard University Press.

Martín-Blas, T., & Serrano-Fernández, A. (2009). The role of new technologies in the learning process: Moodle as a teaching tool in Physics. *Computers & Education*, *52*(1), 35-44.

Peña-López, I. (2010). Are the new millennium learners making the grade. Technology and educational performance in PISA.

Petko, D. (2012). Teachers' pedagogical beliefs and their use of digital media in classrooms: Sharpening the focus of the 'will, skill, tool'model and integrating teachers' constructivist orientations. *Computers & Education*, *58*(4), 1351-1359.

Pérez-Pérez, M., Serrano-Bedia, A. M., & García-Piqueres, G. (2020). An analysis of factors affecting students perceptions of learning outcomes with Moodle. *Journal of Further and Higher Education*, *44*(8), 1114-1129.

Ruggiero, D., & Mong, C. J. (2015). The teacher technology integration experience: Practice and reflection in the classroom. *Journal of Information Technology Education*, *14*.

Vygotsky, L. S. (1978). Socio-cultural theory. Mind in society, 6, 52-58.

Wells, G. (1999). Dialogic inquiry: Towards a socio-cultural practice and theory of education. Cambridge University Press.

Wenger, E. (1998). Communities of practice: Learning as a social system. *Systems thinker*, *9*(5), 2-3.

Williams, M., & Burden, R. (1997). Motivation in language learning: A social constructivist approach. *Cahiers de l'APLIUT*, *16*(3), 19-27.

From Mathematical Thinking to Computational Thinking: Use Scratch Programming to Teach Concepts of Prime and Composite Numbers

Siu Cheung KONG[a*] & Wai Ying KWOK[b]

[a]*Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong*

[b]*Centre for Learning, Teaching and Technology, The Education University of Hong Kong, Hong Kong*

*sckong@eduhk.hk

Abstract: This study pioneered the pedagogical use of Scratch programming to support Grade 6 students to cross from mathematical thinking to computational thinking (CT) in mathematics classrooms. A Scratch-based pedagogical innovation was designed to expose students to the pedagogy "To Play, To Think, To Code" with two Scratch apps and five Scratch activity worksheets to explore, think about, apply and consolidate the target mathematical concepts through Scratch programming. An eight-lesson teaching in 320 minutes was trialed in 15 selected Grade 6 classes involving 324 students from seven primary schools in Hong Kong. From the pre-post-tests, the pedagogical innovation successfully supported students to make statistically significant growth in understanding all five topic-specific mathematical concepts and all five target CT concepts. From the questionnaire surveys, students demonstrated a high level of awareness of the two target CT practices, and a positive perception of CT development for their own good. From the focus group interviews, students confirmed the effectiveness of and expressed a satisfaction with the pedagogy for mathematics learning and CT development through coding. The positive results of this study confirm the potential of the pedagogical innovation which integrates CT education with subject-specific curriculum delivery for an effective development of both subject knowledge and CT competency among primary school students. Implications for the flow and scope of future integration of subject lessons with coding activities are discussed.

Keywords: Computational thinking, mathematical thinking, primary schools, prime and composite numbers, Scratch programming

1. Introduction and Background of Study

Computational thinking (CT) is advocated essential for everyone to succeed in the digitalized society (Grover & Pea, 2013; Shute, Sun, & Asbell-Clarke, 2017). Wing (2006, p.33) defines CT as a thinking process for "solving problems, designing systems, and understanding human behavior, by drawing on the concepts fundamental to computer science". CT has three dimensions: CT concepts – the common concepts used in programming such as sequence, conditionals, repetition; CT practices – the process of programming such as iterative and incremental, abstracting and modularizing, testing and debugging; and CT perspectives – students' understandings of themselves and the technological world (Brennan & Resnick, 2012; Rodríguez-Martínez, González-Calero, & Sáez-López, 2020). Educators realize the necessity to develop CT among students, and so the importance of CT education in school curriculum. The advent of block-based programming environments such as Scratch creates potential to introduce CT activities in subject teaching at primary grades for co-developing subject knowledge and CT (Gadanidis, 2015; Lee, Grover, Martin, Pillai, & Malyn-Smith, 2020). This study tapped into such pedagogical potential – to pioneer the use of Scratch programming among senior primary students to co-develop mathematical concepts of "Prime and Composite Numbers" and CT competency.

The "Number" strand is a central component in primary mathematics curriculum; and "Prime and Composite Numbers" is a main topic in the "Number" strand (Ustunsoy, Ozdemir, & Unal, 2011;

Zazkis & Zazkis, 2014). This topic has three vital knowledge points – 1) any natural number greater than "1" is either a prime or a composite; 2) when a number is represented as a product, it is a composite number unless the factors are "1" and a prime number; and 3) composite numbers have a unique prime decomposition (Dickerson & Pitman, 2016; Ustunsoy et al., 2011). There is a big challenge among students when they learn this mathematical topic – an inadequacy of the concept that a prime number has exactly two factors, not more and not less (Mohyuddin & Khalil, 2016; Ustunsoy et al., 2011). Researchers such as Dickerson and Pitman (2016) and Zazkis and Zazkis (2014) advocate the best way to master the knowledge points and address the learning challenge abovementioned is to develop the concept of using the number of factors of the given natural numbers for categorization – prime numbers have only two factors; while composite numbers have more than two factors.

The block-based programming environment Scratch is popularly used in subject classrooms in primary education (Benton, Hoyles, Kalas, & Noss, 2017; Rodríguez-Martínez et al., 2020). Its intuitive interface-design allows children to make simple actions on dragging, dropping, and combining code blocks to easily create programs and immediately observe the programming outcomes (Calder, 2019; Gadanidis, 2015). Frameworks by Brennan and Resnick (2012) and Grover et al. (2017) are widely referred for integrating CT education into subject curriculum via Scratch programming environment – wherein the coding products serve as computational manipulatives which conceptually align with the traditional notion of educational manipulatives (Calder, 2019; Rodríguez-Martínez et al., 2020).

There is a natural fit to integrate CT education into mathematics curriculum delivery, due to a shared logical structure in the developmental process of algorithmic thinking between mathematical thinking and CT (Gadanidis, 2015; Pérez, 2018). Algorithmic thinking emphasizes the use of a series of ordered steps to solve problems (Rich, Yadav, & Schwarz, 2019; Yadav, Stephenson, & Hong, 2017). There is a complementary connection in students' development between mathematical thinking and CT – to link up abilities of pattern generalization and abstraction (Pérez, 2018; Rich, Spaepen, Strickland, & Moran, 2020). The ability of pattern generalization sets to analyze algorithmic representations to discover regularities within that set of algorithmic representations. Then, the ability of abstraction sets to translate the discovered regularities into a concise and precise mathematical formula, and create an abstraction realizing mathematical formula to be a programmable solution to solve contextual problems automatically. This developmental process is important yet difficult for young students to achieve.

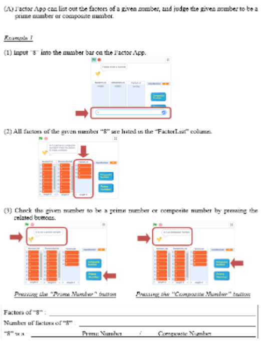
In mathematics subject, there are three main criteria for the pedagogical designs for embedding CT education in subject curriculum. First, the pedagogical designs give students enough chances to work on the selected coding products to construct subject knowledge and stimulate their interest in coding (Calder, 2019; Rodríguez-Martínez et al., 2020). Second, the pedagogical designs give students enough chances to apply subject knowledge in thinking about programming solutions for solving problems in subject-specific contexts (Chiang & Qin, 2018; Rodríguez-Martínez et al., 2020). Third, the pedagogical designs give students enough chances to consolidate subject knowledge in generating coding products for solving subject-specific problems (Benton et al., 2017; Chiang & Qin, 2018).

2. The Study: Research Design and Evaluation Methods

This study pioneered the research on developing both subject knowledge and CT competency through block-based programming activities in subject classrooms. It aimed to innovate a pedagogical design which engages students in Scratch programming for developing the important knowledge of using the number of factors to formulate the concepts of prime numbers (having only two factors) and composite numbers (having more than two factors); and at the same time the competency of five CT concepts ("sequences", "events", "conditionals", "repetition", and "operators"), two CT practices ("iterative and incremental" and "testing and debugging"), and one CT perspective ("ability to connect") – as in the CT frameworks by Brennan and Resnick (2012) and Rodríguez-Martínez et al. (2020).

The pedagogical innovation consisted of a three-step pedagogy "To Play, To Think, To Code" and a Scratch programming environment with two Scratch apps for stimulating students to use the number of factors to formulate the concepts of prime and composite numbers, and so to classify the given numbers to be prime or composite numbers. Five Scratch activity worksheets were also designed to support students to learn using the two Scratch apps. Figure 1 illustrates how students were engaged in "playing" the Factor App for inquiry-based learning of prime and composite numbers; "thinking" of the target concepts through guided-discovery worksheets; and “coding” in Scratch for a programming solution through reusing and refining codes in the Simple Remainder App and the Factor App.

“To Play” step
(Guiding students to explore and observe the fundamental criterion for differentiating prime numbers and composite numbers)



(A) Factor App can list out the factors of a given number, and judge the given number to be a prime number or composite number.

Example 1

(1) Input “8” into the number bar on the Factor App.

(2) All factors of the given number “8” are listed in the “FactorList” column.

(3) Check the given number to be a prime number or composite number by pressing the related buttons.

Pressing the “Prime Number” button *Pressing the “Composite Number” button*

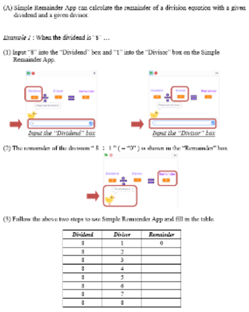
Factors of “8” : ______

Number of factors of “8” : ______

“8” is a : ______ Prime Number / Composite Number ______

Students used the Factor App to explore different numbers for a judgment to be “prime numbers” and “composite numbers”.

“To Think” step
(Guiding students to think about and then generalize the fundamental pattern behind the way to find factors of the given numbers)



(A) Simple Remainder App can calculate the remainder of a division equation with a given dividend and a given divisor.

Example 1 : When the dividend is “8” …

(1) Input “8” into the “Dividend” box and “1” into the “Divisor” box on the Simple Remainder App.

Input the “Dividend” box *Input the “Divisor” box*

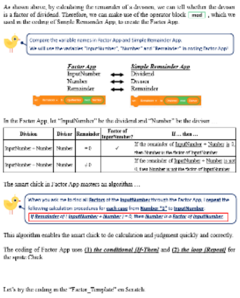
(2) The remainder of the division “ 8 ÷ 1 ” (= “0”) is shown in the “Remainder” box.

(3) Follow the above two steps to use Simple Remainder App and fill in the table.

Dividend	*Divisor*	*Remainder*
8	1	0
8	2	
8	3	
8	4	
8	5	
8	6	
8	7	
8	8	

Students used the Simple Remainder App to calculate and tabulate the remainders of the given division-equations.

“To Code” step
(Guiding students to apply and consolidate the learned knowledge through coding the Simple Remainder App and the Factor App)



As shown above, by calculating the remainder of a division, we can tell whether the divisor is a factor of dividend. Therefore, we can make use of the operator block mod , which we used in the coding of Simple Remainder App, to create the Factor App.

Compare the variable names in Factor App and Simple Remainder App.
We will use the variables “InputNumber”, “Number” and “Remainder” in coding Factor App!

Factor App		*Simple Remainder App*
InputNumber	⟷	Dividend
Number	⟷	Divisor
Remainder	⟷	Remainder

In the Factor App, let “InputNumber” be the dividend and “Number” be the divisor …

Division	Divisor	Remainder	Factor of InputNumber?	If … then …
InputNumber ÷ Number	Number	= 0	✓	If the remainder of InputNumber ÷ Number is 0, then Number is the factor of InputNumber
InputNumber ÷ Number	Number	≠ 0		If the remainder of InputNumber ÷ Number is not 0, then Number is not the factor of InputNumber

The smart chick in Factor App masters an algorithm …

When you ask me to find all Factors of the InputNumber through the Factor App, I repeat the following calculation procedures for each case from Number “1” to InputNumber:
If Remainder of (InputNumber ÷ Number) = 0, then Number is a Factor of InputNumber.

This algorithm enables the smart chick to do calculation and judgment quickly and correctly.

The coding of Factor App uses ***(1) the conditional (If-Then)*** and ***(2) the loop (Repeat)*** for the sprite Chick.

Let’s try the coding in the “Factor_Template” on Scratch.

Students made before-coding reflection on the learned knowledge of the relationship between “Remainders” and “Factors” for finding factors of the given numbers.

Figure 1. Sample Questions in Scratch Activity Worksheets in the “To Play, To Think, To Code” Steps.

This study had 15 Grade 6 classes from seven Hong Kong primary schools – involving a total of 324 students – for a consented participation (see Table 1). The students had some background knowledge of the target mathematical topic, as they learned this topic previously in Grade 4 or Grade 5 mathematics curriculum. The students had some experience in programming before participating in this study. The mathematics teachers of these 15 participating classes trialed the pedagogical innovation on a class-specific basis. Before the trial teaching, the participating teachers completed a four-hour training workshop which prepared them well for a sound recognition of the rationale of learning through coding in local mathematics curriculum; a ready implementation of the “To Play, To Think, To Code” pedagogy in mathematics classroom; and a confident integration of the two Scratch apps and the five Scratch activity worksheets into topic-specific lessons. This study focused on two research questions: (1) What did the students achieve in developing mathematical concepts and CT under the pedagogical innovation? (2) How did the students perceive the pedagogical innovation for developing CT in mathematics classrooms? Three methods were adopted for evaluating the pedagogical innovation, adopting the research instruments developed by the research team with backgrounds in mathematics, education, and computer sciences.

Table 1. *Profile of Students Participated in This Study.*

	School A	School B	School C	School D	School E	School F	School G
No. of students	28	82	68	75	32	12	27
No. of classes	1	3	4	4	1	1	1
Boys : Girls	16:12	37:45	36:32	45:30	17:15	8:4	12:15
Mean age (years)	11.11	10.85	10.93	10.94	10.84	10.92	11.00

Firstly, the pre-post-tests were conducted at the beginning and the end of the pedagogical innovation to investigate students’ achievement in mathematics learning and CT development. The test papers contained 15 questions: four questions on testing the concept of composite and prime numbers, one on the concept of “1” is neither prime number nor composite number, two on the relationship between composite numbers and multiples, three on finding prime numbers by enumeration, one on finding factors of a number, and four on CT concepts including “operator”, “repetition”, “events &

conditionals" and "sequence". A statistical comparison of students' pre-test and post-test scores was conducted with the assistance of SPSS software. The Cronbach's alpha reliability coefficients for the pre-test and post-test are 0.785 and 0.754 respectively.

Secondly, the pre-post-surveys were conducted at the beginning and the end of the pedagogical innovation to investigate students' perception of developing CT in mathematics classrooms. The questionnaire contained five 5-point Likert scale questions, of which three questions on the building of awareness, interest and confidence in programming, and two on the development of CT practices including "iterative and incremental" and "testing and debugging". The mean rating for each question and the corresponding standard deviation were then calculated. The Cronbach's alpha reliability coefficients for the pre-survey and post-survey are 0.827 and 0.861 respectively.

Thirdly, focus group interviews were conducted at the end of the pedagogical innovation to investigate students' perception of the pedagogical innovation for developing CT. A total of 25 students were randomly selected from the seven participating schools, with each focus group consisting of three to five students. The student respondents were asked about how they perceived the help from the pedagogical innovation in their development of mathematical concepts and CT competency, the enjoyment in and satisfaction with the pedagogy for mathematics learning through coding, and the challenges in and recommendations for mathematics lessons integrated with coding activities. All the interview content was transcribed and systematically summarized.

3. Results and Discussion

3.1 Students' Achievement in Developing Mathematical Concepts and CT

The pre-post-tests found that the pedagogical innovation effectively supported students to develop mathematical concepts on prime and composite numbers (see Table 2). The students had a statistically significant increase in the post-test scores for the question items on all topic-specific knowledge points.

Table 2. *Students' Achievement in Developing Concepts of Prime and Composite Numbers Before and After the Pedagogical Innovation (N = 324).*

Question items			Pre-test scores		Post-test scores		
Topic concepts	No. of items	Max. scores	Mean	(SD)	Mean	(SD)	t-test
A. Concept of Factor	7	14	6.91	(3.16)	9.00	(3.17)	12.53***
(1) Concept of composite and prime numbers	4	8	4.45	(2.18)	5.87	(2.12)	11.99***
(2) Concept of "1" is neither prime number nor composite number	1	2	1.18	(0.62)	1.46	(0.57)	6.87***
(3) Exploring the relationship between composite numbers and multiples	2	4	1.28	(1.06)	1.67	(1.17)	5.47***
B. Finding Factors and Prime Numbers	4	10	8.02	(2.37)	8.67	(1.86)	5.51***
(4) Finding prime numbers by enumeration	3	6	4.69	(1.55)	5.18	(1.21)	6.04***
(5) Finding factors of a number	1	4	3.33	(1.17)	3.48	(1.00)	2.20*
C. Total	11	24	14.93	(4.94)	17.66	(4.46)	12.01***

$*p < 0.05$ $***p < 0.001$

For learning the concept of factor, the pre-post-test results indicate that students after the pedagogical innovation had a noticeable knowledge gain – with the dimension-specific mean score below the passing score (i.e. half of the maximum score) in the pre-test to finally above the passing score in the post-test. The students were found to improve greatly their concept of composite and prime numbers, and extend their understanding that "1" is neither prime number nor composite number. For learning the ways of finding factors and prime numbers, students after the pedagogical innovation built on their fairly good knowledge foundation in this dimension for a statistically significant improvement in the knowledge of finding prime numbers by enumeration and finding factors of a number.

The pre-post-tests also found that the pedagogical innovation effectively supported students to develop CT concepts (see Table 3). The students had a statistically significant increase in the post-test scores for the question items on the target CT concepts. The pre-post-test results indicate that students after the pedagogical innovation had a noticeable growth in the mastery of CT concept "operator" – with the mean score below the passing score in the pre-test to finally above the passing score in the post-test. The students were found to maintain their level of mastery of the CT concept "Repetition" throughout the trial teaching – a fair performance of which the mean scores of the related question item in the pre-test and post-test are just-above the passing score, without a statistically significant difference.

Table 3. *Students' Achievement in Developing Concepts of CT Before and After the Pedagogical Innovation (N = 324).*

Question items				Pre-test scores		Post-test scores		
CT concepts		No. of items	Max. scores	Mean	(SD)	Mean	(SD)	t-test
(1)	Operator	1	1	0.45	(0.50)	0.66	(0.48)	5.87***
(2)	Repetition	1	1	0.51	(0.50)	0.50	(0.50)	0.29
(3) & (4)	Events & Conditionals	1	1	0.34	(0.48)	0.44	(0.50)	2.74**
(5)	Sequence	1	1	0.38	(0.49)	0.45	(0.50)	2.27*
Total		4	4	1.68	(1.10)	2.05	(1.16)	5.20***

**p < 0.05 **p < 0.01 ***p < 0.001*

3.2 Students' Perception of Developing CT in Mathematics Classrooms

From Table 4, there is no statistical significance in students' perception of the pedagogical innovation for developing CT in mathematics classrooms before and after the trial teaching. As mentioned, the students had some experience in programming before participating in this study. This possibly led the students to keep a high level of agreement with the importance of the step-by-step development of a program (CT practice of "iterative and incremental") and the operability-testing of the program (CT practice of "testing and debugging") before and after the trial teaching.

Table 4. *Results of Students' Questionnaire Survey on the Perception of the Pedagogical Innovation for Developing CT in Mathematics Classrooms (N = 324).*

Items	Pre-survey		Post-survey		t-test
	Mean (1-5) #	(SD)	Mean (1-5) #	(SD)	
I think it is important to test the program to make sure it works.	4.12	(0.96)	4.04	(1.11)	1.14
I think it is important to develop a program step by step.	4.09	(0.98)	4.00	(1.08)	1.30
I think that programming is important in our daily lives.	3.59	(1.03)	3.63	(1.13)	0.64
I am confident that I can write a simple program.	3.38	(1.15)	3.39	(1.17)	0.14
I am interested in learning programming.	3.35	(1.13)	3.32	(1.15)	0.55

#*Note: 1 = "strongly disagree", 2 = "disagree", 3 = "neutral"; 4 = "agree"; 5 = "strongly agree".*

The focus group interviews further confirmed students' positive perception of the pedagogical innovation for developing CT in mathematics classrooms (see Table 5). Echoing with the results of pre-post-tests and surveys, the students confirmed that the mathematically-rich activities on using Scratch apps can effectively support them to master the mathematical concepts on prime and composite numbers, as well as the CT concepts and CT practices targeted at this study. All student respondents indicated that the trial teaching enabled them to master the fundamental concept that when an integer can divide a number without giving a remainder, that integer is the "factor" of the number being divided. A student respondent illustrated an example that for the number “10”, the multiplication equations “1 x 10 = 10” and “2 x 5 = 10” stand; and so the number “10” has “1”, “2”, “5” and “10” being its four factors. He further elaborated that this number has more than two factors; and so by definition it is a composite number but not a prime number which has two factors only. Nearly a fifth of the student respondents expressed that “1” is an important knowledge point in the trial teaching – they were able to explicate that “1” has only one factor, the number itself; and so this number does not meet the definition of neither a “prime number” nor a “composite number”. They pointed out that during the trial teaching many students, without the instruction or request from the teachers, tried to use the apps to check the category of “1” and discover this uniqueness of “1”, and in turn developing the correct understanding that “1” is neither a “prime number” nor a “composite number”. Two student respondents indicated that they used the apps to extend their learning exploration of the number “0” which is seldomly covered by traditional mathematics textbook; and noticed that “0” is a special number which neither prime nor composite. This finding implies the need for teachers to discuss with students about the uniqueness of the special numbers “1” and “0”. Nearly three quarters of the student respondents indicated that the trial teaching, comparing with the typical teaching approach, can foster them to think more about the mathematical concepts behind the process of classifying prime and composite numbers through checking the number of factors of the given numbers. These student respondents pointed out that they seldomly think about the meaning and purpose of each calculation step. They appreciated the coding activities dissected each calculation step to give a clear visualization of the calculation process; and this fostered them to widen their thinking angles to look into what and why each of those calculation steps is necessary.

Table 5. *Feedback from Students’ Focus Group Interviews on the Perception of the Pedagogical Innovation for Developing CT (N = 25).*

Major interview feedback
Help in the development of mathematical concepts and CT
• The apps in the trial teaching supported students to quickly identify all factors of the given numbers and accurately classify the given numbers into prime numbers and composite numbers. • The trial teaching enabled students to explore some very large numbers and the special numbers “1” and “0” that are seldomly covered by traditional mathematics textbook, as the apps were convenient to use for finding factors of the given numbers quickly and accurately. • The apps in the trial teaching impressively saved students’ time to manually find factors of the given numbers; and more lesson time can be arranged for student-student and student-teacher interactions to exchange topic-specific concepts. • The steps in the coding activities served as a clear guidance of classifying primes and composites through checking the number of factors of the given numbers. Students were well supported to better understand the concepts behind the process of identifying primes and composites. • The trial teaching motivated students’ extra efforts to refine and debug the codes for the apps after class time to improve existing features and add new features, such as to make the apps able to process negative integers. Students got a great sense of achievement and confidence when the coding products can operate as intended. • This learning experience inspired students to make the best efforts to try different alternatives for solving problems in coding and daily life. Students made many errors at the beginning of the coding process; and they were willing and committed to carefully check the coding outcomes, review their codes, correct the sequence, and vary the parameters of the command blocks for many trials to get the intended coding outcomes.
Enjoyment in and satisfaction with the pedagogy for mathematics learning through coding
• Learning mathematics through coding is considered very interesting and meaningful. • The coding tasks were the most popular part during the trial mathematics lessons. • The apps in the trial teaching were so attractive and interesting to stimulate learning motivation. • There was a great enjoyment in the coding activities which are new and challenging. • Coding tasks were highly satisfied, with a great sense of achievement after coding successfully. • The steps in the activity worksheets were confirmed to be clearly and comprehensively stated. The questions inside were considered effective to guide students to progressively deepen their knowledge of the target topic as well as coding.

Challenges in and recommendations for mathematics lessons integrated with coding activities
• Some technical problems occurred in the coding lessons, as not every student was familiar with Scratch programming. • Learning diversity existed among students in the coding lessons - some students mastered coding quickly while some not. • Each student should have two computing devices during the trial lessons: one for viewing teachers' lecturing, and the other one for coding along with teachers' demonstration. • Pre-training of Scratch coding should be arranged before the trial lessons for familiarization. • Other mathematical topics suitable for learning through coding include "Percentage", "3D Shapes", "Addition and Subtraction", and "Equations". • The approach of learning through coding can be extended to students at lower grades and to other subjects, such as the topic of "Storytelling" in language subjects at Grade 1; the learning of vocabulary in English Language subject; and the learning of the classification of animals in General Studies subject.

Nearly all student respondents appreciated that the coding activities can guide them to discover a series of calculation steps that teachers seldomly mention in typical mathematics classroom when teaching the target topic. This allowed them to develop mathematical concepts and coding knowledge at the same time. A student respondent further indicated his impression that the coding activities equipped them with knowledge of some operators in Scratch coding, such as the [mod] operator-block (i.e. "modulo") for division equations. Nearly 45% of the student respondents demonstrated a high awareness of debugging in the coding process, as reflected in the response that it is a must to order codes in a correct sequence during the coding process, for ensuring the apps are able to operate as intended. Four of them reflected that there were many errors made at the beginning of the coding process, and they were willing and committed to carefully check the coding outcomes, review their codes, correct the sequence, and vary the parameters of the command blocks for many trials to get the intended coding outcomes. They impressively linked the debugging step in coding with the procedures-check in mathematical calculation – that both needed to be attentive to the minute steps in order to give the intended outcomes. This learning experience inspired them to make the best efforts to try different alternatives for solving problems in coding and daily life.

The students also enjoyed and felt satisfied with the pedagogy for mathematics learning through coding. Nearly all student respondents indicated that they liked and committed to use the apps for a quick check on the number of all factors for the given numbers. They preferred to have this innovative approach in mathematics classrooms. Around 45% of the student respondents confirmed that the learning approach in the trial teaching can effectively combine mathematics learning with CT development. They were satisfied with the benefits of developing both mathematical concepts and CT competency in one subject lesson. Nearly 30% of the student respondents confirmed that it was interesting to use the apps for supporting the learning of the target topic. They considered lessons in the trial teaching were less boring, comparing the traditional mathematics classrooms, as they were allowed to play apps during the learning process. Two of them further indicated their extra efforts to refine and debug the codes for the apps after class time for improving existing and add new features, such as to make the apps able to process negative integers. They got a great sense of achievement and confidence when the coding products can operate as intended.

There was feedback on two challenges and four recommendations for mathematics lessons integrated with coding activities. Two student respondents explained that it was their first time to do Scratch coding and they were unfamiliar with Scratch programming environment. This led to two main challenges in the pedagogical innovation: some technical problems occurred in the coding lessons; and some students lagged behind in the progress of coding activities. In this regard, the student respondents indicated their expectation for pre-training of Scratch coding before the trial lessons, followed by the existing appropriate approach of first teaching about the basic concepts of the target mathematical topic, and then teaching about the knowledge and skills of coding. They stressed that it is good for them to have two computing devices during the trial lessons, in which they use one device to view teachers' lecturing and the other device to code along with teachers' demonstration. Four student respondents suggested that "Percentage", "3D Shapes", "Addition and Subtraction", and "Equations" are mathematical topics suitable for learning through coding. The other three student respondents further suggested that it is possible to extend the approach of learning through coding to support students at lower grades; or to support the learning and teaching in other subject topics. One of these student respondents made suggestion on the topic of "Storytelling" in language subjects at Grade 1. The other student respondent made suggestion on the learning of vocabulary in English Language subject; and the learning of the classification of animals in General Studies subject.

3.3 Interaction between Mathematical Thinking and Computational Thinking in Student Learning under the Pedagogical Innovation

From the results of this study, the pedagogical innovation with the three-step "To Play, To Think, To Code" approach can benefit students in the learning of mathematical and computational domains.

The first half of the pedagogical innovation – the mathematically-rich activities on using Scratch apps – exposes students to the interaction from mathematical thinking to CT. This part is attested to be effective, as reflected by students' feedback from focus group interviews on appreciating the support from Scratch apps for them to dissect and visualize each calculation step of finding factors; illustrating the procedures and justification for finding factors of a number such as "10"; as well as indicating the extension of self-initiative to explore the categories and discover the uniqueness of the special numbers "1" and "0". These results also imply that the mathematically-rich activities in the pedagogical innovation are potential to help students address their common topic-specific learning challenge as suggested by Mohyuddin and Khalil (2016) and Ustunsoy et al. (2011) – as students under the pedagogical innovation can understand "1" is not a prime number and firmly grasp the concept that a prime number has exactly two factors.

The second half of the pedagogical innovation – the computationally-rich activities on programming Scratch apps – exposes students to the interaction from CT to mathematical thinking. This part is also attested to be effective, with evidence from focus group interviews in which students explicated the grasp of knowledge about Scratch coding blocks for mathematical operations such as [mod] operator-block in division equations; indicated the awareness of outcomes check, codes review, sequence correctness, and parameters variation during coding process; as well as linked the debugging step in coding with procedure-check in mathematical calculation. These results also imply that the computationally-rich activities in the pedagogical innovation echo with the advocacy from Gadanidis (2015) and Pérez (2018) to be a potential approach to a fitting integration of CT education into mathematics curriculum delivery through linking up the developmental processes of algorithmic thinking between mathematical thinking and CT.

The interaction from mathematical thinking to CT achieved by the students can be attributed to the four types of engagement in the mathematically-rich activities. In the first half of the pedagogical innovation, students first developed *(i) mathematical thinking* through working with Scratch activity worksheets to tabulate and observe the pattern that "If Remainder of Dividend ÷ Divisor is / is not 0, then Divisor is / is not a Factor of Dividend" among the given division-equations. Students based on the tabulation results to generalize the pattern that when a given number is divided by a divisor in the range of 1 to the given number, a "zero" remainder in the related division-equation means the divisor is the factor of that given number. Next, students developed *(ii) computational thinking* to think about the possibility to write a computing program to automatically solve the problem of finding factors of a given number; and the need for the computing program to get a feature to store the factors of the given number. Students were subsequently guided to think about the need to come up with an algorithm to find out all factors of a given number. Accordingly, students were led to link up mathematical thinking with CT through the step of *(iii) abstraction* (i.e. making an abstraction of the mathematical pattern that when a given number is divided by a divisor in the range of 1 to the given number, a "zero" remainder in the related division-equation means the divisor is the factor of that given number); and the step of *(iv) algorithmic thinking* (i.e. setting the algorithm with the necessary variables for the computing program, based on their abstraction of the tabulated mathematical pattern of "If Remainder of Dividend ÷ Divisor is / is not 0, then Divisor is / is not a Factor of Dividend").

The interaction from CT to mathematical thinking achieved by the students can be attributed to the three types of engagement in the computationally-rich activities. In the second half of the pedagogical innovation, students progressively developed *(v) CT concepts* and *(vi) CT practices* through implementing the algorithm in the Scratch programming environment by the step of reusing the codes from the Simple Remainder App to refine the Factor App; and the step of testing if their Scratch program can correctly list out all factors of the given numbers and judge the given numbers as prime numbers or composite numbers. Students' development of CT concepts covered "sequences", "events" [When …], "conditionals [If-Then] and [If-Then-Else]", "repetition [Repeat]", "operators [Mod], [=], [>] and [<]"; and their development of CT practice covered "reusing and remixing", "iterative and incremental" and "testing and debugging". Finally, students were led to develop *(vii) CT perspectives* through connecting the mathematical task on decomposing relatively large prime numbers with the significant role of prime numbers in information security in digital communication – this step demonstrated the CT perspectives of "ability to connect".

The mathematically-rich activities and the computationally-rich activities in the pedagogical innovation on the whole support students on progressing to an interaction between mathematical thinking and CT through linking up the abilities of pattern generalization and abstraction. The series of "To Code" activities started at leading students to develop and demonstrate CT competency through Scratch programming; and ended with fostering students to apply and consolidate their mathematical understanding developed through the "To Play" and "To Think" activities. Apart from these activities, teachers should reflect with students about the uniqueness of the numbers "1" and "0" on top of natural numbers greater than "1" in the context of prime and composite numbers (Kong, 2019).

4. Conclusion and Future Direction

This study developed and implemented a pedagogical innovation which used Scratch programming to support 324 students from 15 selected Grade 6 classes in seven primary schools at Hong Kong to co-develop mathematical concepts and CT competency in subject classrooms. Under an eight-lesson teaching supported by the pedagogy "To Play, To Think, To Code" with two Scratch apps and five Scratch activity worksheets, students in this Scratch-based pedagogical innovation explored, thought about, applied and consolidated the mathematical concepts of prime and composite numbers through Scratch programming; in which also went through the development and application of CT concepts, CT practices, and CT perspectives. The pre-post-tests confirmed the effective support from the pedagogical innovation for students to significantly enhance their understanding of mathematical knowledge about the concepts of factors, composite numbers and prime numbers, and the ways of finding factors and prime numbers; as well as CT concepts of "operator", "events & conditionals" and "sequence". The questionnaire surveys confirmed that the pedagogical innovation successfully fostered students to be highly aware of CT practices of "iterative and incremental" and "testing and debugging". The focus group interviews confirmed that students positively perceived the pedagogical innovation to be an effective and satisfying pedagogy for mathematics learning and CT development through coding.

The evidence found in this study implies that the pedagogical innovation is potential to effectively support primary school students to co-develop mathematical concepts and CT competency. It is promising to first engage students in mathematically-rich activities on using Scratch apps to explore the concepts of prime and composite numbers and generalize the pattern of finding factors of the given numbers; and then guide students to cross from mathematical thinking to CT for making abstraction and algorithmic thinking – to translate the generalized pattern into a mathematical formula in pseudocodes – for a programmable solution which automatically finds factors of the given numbers and categorizes prime and composite numbers; and finally engage students in the computationally-rich activities on programming Scratch apps – in which they develop and apply CT concepts, CT practices, and CT perspectives when generating a programmable outcome for the target solution-automation.

The recommendations collected from this study imply the potential to expand the application scope of the pedagogical innovation to other mathematical topics, other grades, and/or other subjects. One of the possible future directions will be the pedagogical innovation for using Scratch programming in Grade 4 English Language classrooms to support students to develop building blocks for the learning and teaching of locations and directions. The future research will try to address the need of two computing devices for each student to view lecturing and perform coding, for a smoother lesson flow. This study had a limitation of no control group involved for evaluating the effectiveness of the pedagogical innovation. Future research will also try to arrange a control group for evaluation purposes.

Acknowledgements

The authors would like to acknowledge the support of the project “Use Coding as a Pedagogy for Teaching Subject Knowledge in Mathematics, General Studies and English Language and Fostering Computational Thinking” funded by the HKSAR Quality Education Fund.

References

Benton, L., Hoyles, C., Kalas, I., & Noss, R. (2017). Bridging primary programming and mathematics: Some findings of design research in England. *Digital Experiences in Mathematics Education*, *3*(2), 115-138.

Brennan, K., & Resnick, M. (2012). New frameworks for studying and assessing the development of computational thinking. *2012 Annual Meeting of the American Educational Research Association (AERA'12)*, Canada.

Calder, N. (2019). Using Scratch to facilitate mathematical thinking. *Waikato Journal of Education*, *23*(2), 43-58.

Chiang, F.-K., & Qin, L. (2018). A pilot study to assess the impacts of game-based construction learning, using Scratch, on students' multi-step equation-solving performance. *Interactive Learning Environments*, *26*(6), 803-814.

Dickerson, D. S., & Pitman, D. J. (2016). An examination of college mathematics majors' understandings of their own written definitions. *Journal of Mathematical Behavior*, *41*, 1-9.

Gadanidis, G. (2015). Coding as a Trojan Horse for mathematics education reform. *Journal of Computers in Mathematics and Science Teaching*, *34*(2), 155-173.

Grover, S., & Pea, R. (2013). Computational thinking in K-12: A review of the state of the field. *Educational Researcher*, *42*(1), 38-43.

Grover, S., Basu, S., Bienkowski, M., Eagle, M., Diana, N., & Stamper, J. (2017). A framework for using hypothesis-driven approaches to support data-driven learning analytics in measuring computational thinking in block-based programming environments. *ACM Transactions on Computing Education*, *17*(3), 14.

Kong, S. C. (2019). Learning composite and prime numbers through developing an app: An example of computational thinking development through primary mathematics learning. In S. C. Kong & H. Abelson (Eds.), *Computational thinking education* (pp. 145-166). Singapore: SpringerOpen.

Lee, I., Grover, S., Martin, F., Pillai, S., & Malyn-Smith, J. (2020). Computational thinking from a disciplinary perspective: Integrating computational thinking in K-12 science, technology, engineering, and mathematics education. *Journal of Science Education and Technology*, *29*, 1-8.

Mohyuddin, R. G., & Khalil, U. (2016). Misconceptions of students in learning mathematics at primary level. *Bulletin of Education and Research*, *38*(1), 133-162.

Pérez, A. (2018). A framework for computational thinking dispositions in mathematics education. *Journal for Research in Mathematics Education*, *49*(4), 424-461.

Polly, D. (2011). Examining how the enactment of TPACK varies across grade levels in mathematics. *Journal of Computers in Mathematics and Science Teaching*, *30*(1), 37-59.

Rich, K. M., Spaepen, E., Strickland, C., & Moran, C. (2020). Synergies and differences in mathematical and computational thinking: Implications for integrated instruction. *Interactive Learning Environments*, *28*(3), 272-283.

Rich, K. M., Yadav, A., & Schwarz, C. V. (2019). Computational thinking, mathematics, and science: Elementary teachers' perspectives on integration. *Journal of Technology and Teacher Education*, *27*(2), 165-205.

Rodríguez-Martínez, J. A., González-Calero, J. A., & Sáez-López, J. M. (2020). Computational thinking and mathematics using Scratch: An experiment with sixth-grade students. *Interactive Learning Environments*, *28*(3), 316-327.

Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review*, *22*, 142-158.

Ustunsoy, S., Ozdemir, A. S., & Unal, H. (2011). The investigation of student approach to problem solving about some topics of Number Theory. *Procedia - Social and Behavioral Sciences*, *15*, 3422-3425.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33-35.

Yadav, A., Stephenson, C., & Hong, H. (2017). Computational thinking for teacher education. *Communications of the ACM*, *60*(4), 55-62.

Zazkis, R., & Zazkis, D. (2014). Script writing in the mathematics classroom: Imaginary conversations on the structure of numbers. *Research in Mathematics Education*, *16*(1), 54-70.

Mining Students' Engagement Pattern in Summer Vacation Assignment

Hiroyuki KUROMIYA[a*], Rwitajit MAJUMDAR[b] & Hiroaki OGATA[b]
[a]*Graduate School of Informatics, Kyoto University, Japan*
[b]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*khiroyuki1993@gmail.com

Abstract: Learning Analytics (LA) is an emergent field which aims at a better understanding of students and providing intelligence to learners, teachers, and administrators using learning log data. Although the use of technology in class is increasing in the K-12 sector as well as territory education, cases of effective implementation of LA in secondary schools were rarely reported, especially in Japan. In this paper, we offer an example where LA is implemented at a junior-high Math class in Japan. We introduce our LA platform, LEAF - LMS and e-book integrated learning analytics dashboard - and its usage during summer vacation period in the target class. We analyzed 121 students' question answering logs and their exam performance after the vacation by K-means clustering method. As a result, we found that students' progress patterns were able to be categorized as four types: early engagement, late engagement, high engagement, and low engagement and the early and high engagement group got significantly higher scores than the low engagement group. It implies the importance of the engagement at the beginning of the vacation. Moreover, by comparing the previous studies in MOOCs, we concluded that self-regulation skills are an important factor for student success in a long vacation period, too. Finally, we introduce a monitoring tool which aims to detect and send messages to at-risk students at an early stage in the next summer vacation period. Our case will become the first model case of how to implement LA in secondary school in Japan.

Keywords: Learning Analytics, Secondary Education, Long Vacation Period, Pattern Mining

1. Introduction

Learning Analytics (LA) is an emergent field which aims at a better understanding of students and providing intelligence to learners, teachers, and administrators using learning log data (Law & Liang, 2020). Higher Education (HE) has been using LA for improving the services and students' retention rate (Bienkowski et al., 2012). However, adopting LA in school is not an easy task. It is estimated that it will take two or three years to adopt LA within primary and secondary education (Freeman et al., 2017). There are many barriers for the adoption of LA in K-12 context: the privacy issues are more sensitive (Gunawardena, 2017), resources for supporting analytics implementation more constrained, and expertise in data analytics is very limited in school context (Kovanović et al., 2020). As a result, it is reported that the number of studies conducted in schools is much fewer than that of in higher institutes (Li et al., 2015); 82.9 % of studies focus on higher education while 17.1 % of secondary school. As a matter of course, cases from Japanese junior-high schools that report LA implementations in school were very limited in the current state.

In this paper, we offer an example where LA is implemented at a junior-high Math class in Japan. In particular, we focused on a summer vacation period in the target school. We addressed the following problems in this paper: 1) What kind of learning patterns were extracted from students' log data during a long vacation period? and 2) How were the extracted patterns related to their performance? Our paper is structured as follows. Related works introduces research on learning analytics implementations at school level education and the prior studies about students' engagement patterns. In the Methodology section, we describe our implementations of LA in the target school and the study settings where we conducted in a junior-high Math class in Japan. Results section shows typical students' learning patterns in summer vacation period extracted by an unsupervised clustering analysis and the relationship with their academic performance after the vacation. Discussion section summarizes our findings and proposes a development of a system which enables timely-intervention in the next summer vacation period for us.

2. Literature Review

2.1 Learning Analytics Implementation at School

As we have mentioned above, there are not so many cases from secondary education context in the learning analytics field. Although the use of technology in class is increasing in the K-12 sector (Horn & Staker, 2011), just using technology in classrooms is not enough for what we call LA. According to Clow (2012), LA is defined as a cycle which consists of four phases - learners, data, metrics, and intervention. Closing the loop is a crucial factor for successful implementation of LA (Corrin et al., 2020). In that sense, the cases about the implementation of LA in schools are very limited in the current state.

Here, we introduce some examples and trends about the adoption of LA in school level education. In Spanish, a research project called PILARES (Smart Learning Analytics Platform to enhance Performance in Secondary Education) was developed for blended learning in secondary school (Sancho et al., 2015) financed by the Spanish government and the collaboration of the Catalan Ministry of Education. It includes a large Moodle based LMS called AGORA, that is used by more than 1,500 schools in Catalonia, and it aims at building a LA platform to allow better insight of the learning process through the LMS. In Uruguay, a countrywide LA tool was introduced for secondary education (Macarini et al., 2019). Although they shared several challenges and constraints they faced during its conception and development, they pointed out the feasibility of finding meaningful patterns using the data obtained from the database, and proposed a prototype for tracking the students' scholar trajectory. Although the substantial growth of the learning analytics field itself provided more possibilities to use LA in primary and secondary education (Ochoa et al., 2017), the actual implementations are very limited in K-12 context.

2.2 Student Engagement Patterns

Students' engagement patterns have been investigated mainly using psychological questionnaires to students. Schnitzler et al. (2020) analyzed 397 high school students' profiles by latent profile analysis (LPA) based on three indicators - participation, cognitive engagement, and emotional engagement. Although the first indicator was assessed by the number of hand-raisings in the classroom, the others were measured with survey items. Finally, they discovered five engagement patterns - disengaged, compliant, silent, engaged, and busy - and the significant differences among the learning patterns. The similar approach was taken in the U.S., targeting 1125 middle school students in a science course to identify engagement profiles and their relationship with science achievement (Bae & DeBusk-Lane, 2019) . By applying LPA, they discovered five engagement types - Moderately engaged/disengaged, Behaviorally engaged/disengaged, and Disengaged - from the survey of students.

Some studies used learning log data to categorize students' engagement patterns. Ebook reading logs were used to categorize students' study patterns at a university in Japan (Akçapinar et al., 2020). They constructed study sequences based on the timestamp they opened the material from the click-stream data and adopted hierarchical cluster analysis to the dataset. As a result, they found three different study patterns from the dataset. MOOCs' interaction logs with lectures and assignments were also used to identify learners' study patterns (Boroujeni & Dillenbourg, 2018). They extracted the action sequences from learners' log data and transformed them into probability distribution matrices for distance computing. By the analysis, they classified students into mainly two types - fixed approach and changing approach - based on their strategy of learning. In the analysis, both hypothesis-driven and data-driven approaches were taken in this study.

2.3 Research Questions

As we have seen before, the LA implementations in secondary schools were very limited; at the same time, student engagement study during a long vacation period in schools were very few. In this paper, we investigate following two questions regarding the LA utilization at school during a long vacation period. In Japanese secondary school context, long vacation period is not only a break from study, but also home study period for students. Usually, students are assumed to engage in their home working assignments given by the teacher before the vacation period.

RQ1: What kind of learning patterns were extracted from students' log data during a long vacation period?

Here, we investigate the study patterns of students in a junior-high school during a long vacation period using the LA platform implemented at the school. Standard clustering techniques based on the timestamp of the engagement from students' interaction logs with an ebook reader will be used to answer this question.

RQ2: How were the extracted patterns related to their academic performance?

To answer this question, we explore the relationship between students' study patterns extracted in the previous question and their academic performance after the vacation. Statistical testing of average exam score by each cluster will be conducted to check if there are any significant differences among clusters.

In the next section, we described a case of a junior-high school in Japan during a summer vacation period. We introduce a specific learning analytics platform called LEAF and offer a use case of it during a summer vacation period in the target school. Then, we analyzed the data retrieved from the platform and discussed the possibility of it.

3. LA Implementations for School in Our Context

3.1 LEAF Platform

Since April 2019, we have been offering an LMS-integrated learning analytics platform called LEAF (Learning Evidence Analytics Framework) to a junior-high school in Japan (see Figure 1). LEAF has three online learning tools - Moodle, BookRoll, and LAViEW (Flanagan & Ogata, 2018). In LEAF, school teachers upload learning materials to an ebook reader BookRoll, and students view the materials as they need. Teachers are also able to post quizzes or recommendations (external link) to their materials and students can comment, highlight text, and write handwritten memos on that material. BookRoll is accessed from Moodle LMS by LTI (Learning Tools Interoperability) authentication method, so users can log in to BookRoll without creating an account for it. It gives a benefit to researchers as well because we can retrieve students' information by their moodle ids. The last component, LAViEW is a dashboard that visualizes the learner-content interactions in BookRoll. Teachers can see students' highlights, memos, and time spent on each page.

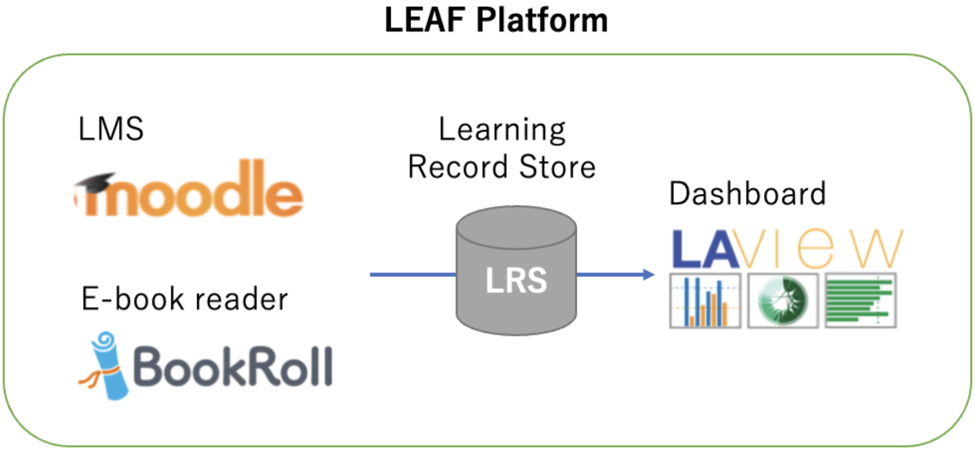
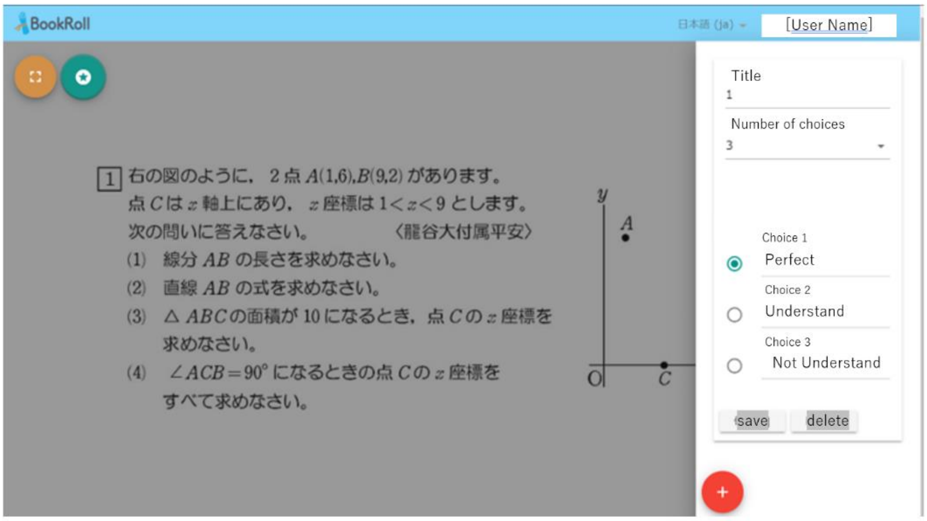


Figure 1. Three Learning Tools in the LEAF Platform in the Target School.

3.2 *Use Case Scenario: Monitoring Summer Vacation Assignment with LEAF*

In this paper, we focus on a specific use case scenario of the LEAF platform in summer vacation period in Math class. We targeted three classes containing 121 students in a junior-high third grade Math course. Before the summer vacation, a teacher uploaded an assignment containing forty-nine Math questions to BookRoll (see Figure 2). The assignment consists of one question per page, and one simple questionnaire is implemented per page. The questionnaire has options which represent three different understanding levels - perfect, understood, or not well understood - as you can see from Figure 2 on the right hand side. In the summer vacation period, students are required to answer the questionnaire each and every time students finish solving a problem as well as submit a paper which contains the answers and working formulas to all the questions. The duration of the summer vacation was from 22 July to 21 August 2019. To measure students' performance, the examination was conducted on 23 August after the summer vacation. We used this score to investigate the relationship between students' behavior and performance.



1 右の図のように，２点 $A(1,6),B(9,2)$ があります。
点 C は x 軸上にあり，x 座標は $1<x<9$ とします。
次の問いに答えなさい。〈龍谷大付属平安〉
(1) 線分 AB の長さを求めなさい。
(2) 直線 AB の式を求めなさい。
(3) $\triangle ABC$ の面積が 10 になるとき，点 C の x 座標を求めなさい。
(4) $\angle ACB=90^\circ$ になるときの点 C の x 座標をすべて求めなさい。

Title
1
Number of choices
3
Choice 1
● Perfect
Choice 2
○ Understand
Choice 3
○ Not Understand

Figure 2. Summer Vacation Assignment Broadcasted via Bookroll.

4. Analysis Procedure

4.1 *Preprocessing of the E-book Logs*

To measure students' engagement during the summer vacation period, we used the students' answer logs to the questionnaire on the assignment. By analyzing students' answer logs, we can get which students solved which question on what day. For the log analysis, we targeted ebook logs from 17 July to 22 August. Overall, we extracted 6,250 answers from the students. we excluded 1,172 answers from page no.1 where the content was just a description of the assignment. Then, we excluded 132 answers after the exam. As a result, we analyzed 4,936 answers by students. Throughout this process, sixteen students were excluded because they had no answer data. Moreover, three students were excluded because they had no score data. In total, we got 102 students.

4.2 *Categorization of the Learning Patterns*

As the summer vacation period was over thirty-seven days, we separated the answers into two periods: logs before 4 August were from 'first-half' and after were from 'second-half.' Duplicated answers to the same question in the same period were excluded in this process. Finally, standard K-means was conducted based on the number of interactions during the first-half and the second-half period for each student. The number of clusters was determined by the Elbow plot based on several cluster indicators (AIC, BIC, and WSS).

4.3 Relationship with Performance

After the categorization step, we compared the average exam score after the vacation period among each cluster. ANOVA and Post-Hoc testing were adopted using statistical testing software, JASP (Love et al., 2019).

5. Results

5.1 Clustering Students' Behavior

Figure 3 shows the Elbow plot which represents the transition of three cluster information over nine cluster numbers (from two to ten). Three cluster information is AIC (Akaike Information Criteria), BIC (Bayesian Information Criteria), and WSS (Within Sum of Squares). Based on the value of BIC, we decided the optimal cluster number as four because it was the lowest BIC value in the plot.

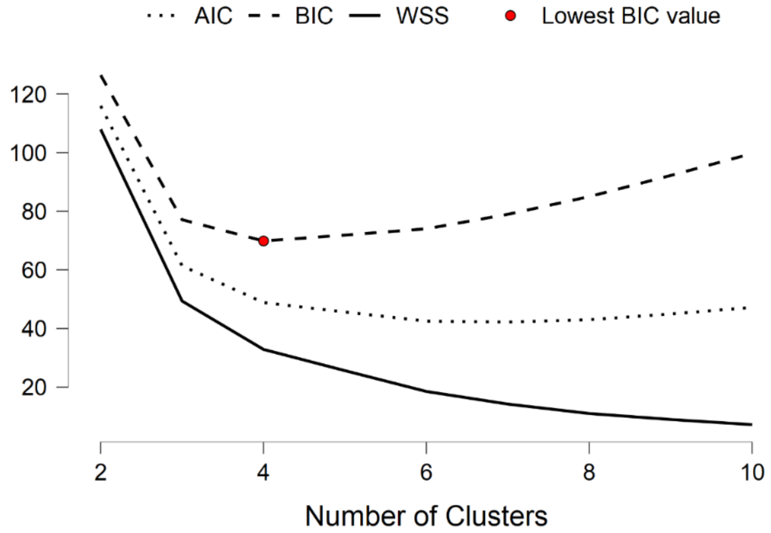


Figure 3. Elbow Plot for Determining Optimal Cluster Number.

Figure 4 shows clustering results when the number of clusters was four. We labeled the cluster based on the location of the scatter plot. Nineteen students were categorized as the early engagement pattern, thirty-two students as high engagement pattern, fourteen students as late engagement pattern, and thirty-seven students as low engagement pattern.

Figure 5 shows actual students' progress patterns for each cluster. Horizontal axis represents the day of the vacation and the vertical axis represents the page of the question they answered. As you can see, students in the early engagement group finished the assignments within the first-half period. On the other hand, students in the late engagement group made little progress by the half of the vacation. Students in the high engagement group can be categorized as two sub groups: some students finished the assignment once by the half of the vacation and solved the questions again at the end of the vacation and others continuously were working throughout the vacation. On the other hand, students in the low engagement group were not able to finish the assignment.

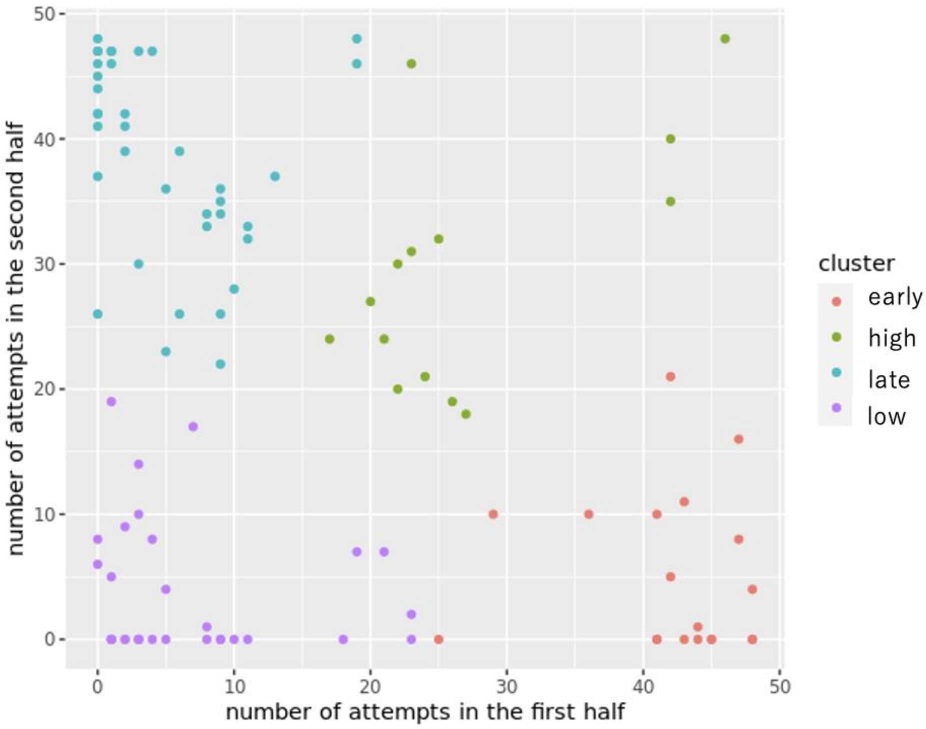


Figure 4. Clustering Results based on the Optimal Cluster Number (K=4).

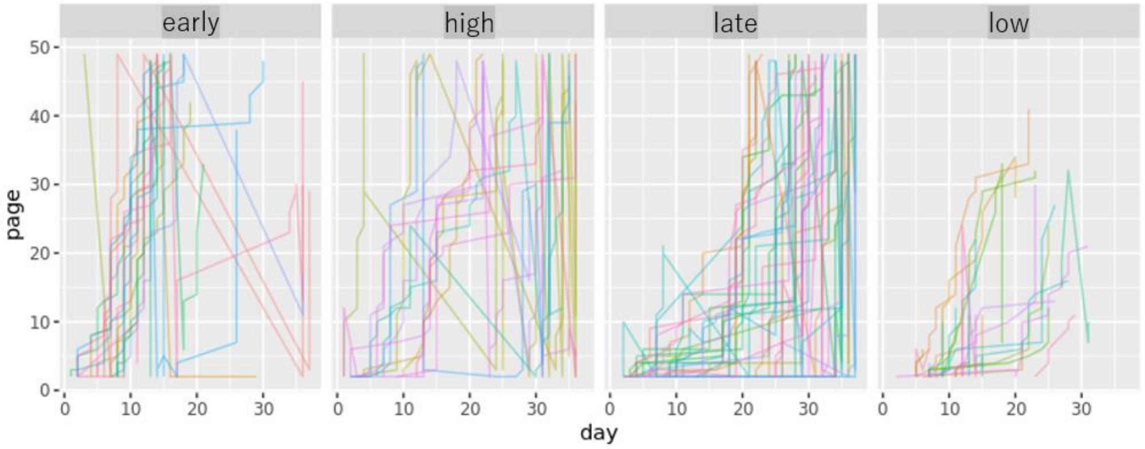


Figure 5. Students' Progress Patterns for each Cluster.

5.2 Relationship with Performance

Finally, we investigated the relationship between their solving patterns and performance. Figure 6 shows the descriptive plot of students' exam scores among four clusters. The error bar stands for standard error of each cluster. As you can see, students in the early engagement group got the highest score (M: 80.8, SD: 14.5), and students in the high engagement group got the second highest score (M: 80.7, SD: 15.0). Students in the late engagement group were the third rank (M: 73.4, SD: 19.9) and students in the low engagement got the lowest score among all (M: 61.9, SD: 21.0). Finally, we conducted ANOVA to test the difference in exam score by clusters. Before adopting ANOVA, we checked the homogeneity of the data and the result was not significant (p = .18). The result of ANOVA was significant (p = .001), so we conducted a Post-hoc comparison between clusters (see Table 1). As a result, we got significant differences between cluster 1 and 4; and cluster 2 and 4. The p-values were adjusted for multiple comparisons.

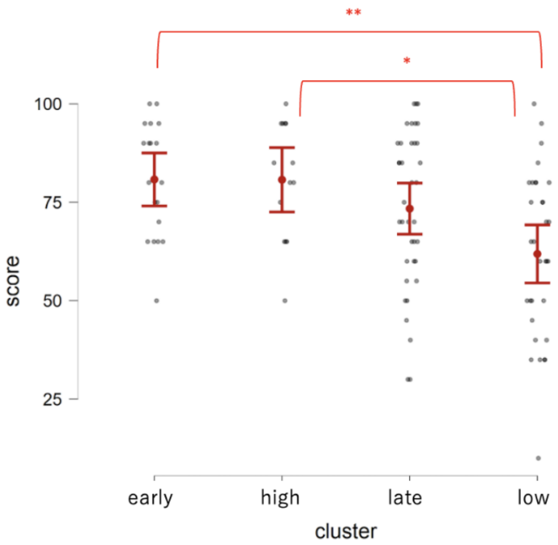


Figure 6. Average Exam Score after the Vacation for each Cluster.

Table 1. *Post-Hoc Comparison Results between Every Cluster Combination*

		Mean Difference	SE	T	P
Early	High	0.075	6.618	0.011	1.000
	Late	7.411	5.303	1.398	0.504
	Low	18.914	5.442	3.476	0.004**
High	Late	7.336	5.896	1.244	0.600
	Low	18.839	6.021	3.129	0.012*
Late	Low	11.503	4.536	2.536	0.061

*Note. p-value adjusted for comparing a family of four (*p<.05, **p<.01)*

6. Discussions and Conclusions

6.1 Summary of Our Findings

By the K-means clustering of the students’ answer logs, we found that students’ progress patterns were able to be categorized as four types:

1. Early engagement: Students who engaged at the beginning of the vacation
2. High engagement: Students who were continuously working throughout the vacation
3. Late engagement: Students who did little until the end of the vacation
4. Low engagement: Students who didn’t finish the assignment

Before our analysis, the teachers in the target school had an assumption that students can be categorized as four types: 1) constant group, 2) early engagement group, 3) late engagement group, and 4) early + review group. Constant engagement group refers to students who continuously work the assignment (e.g. one question per one day) and early + review group refers to students who finished early and review the questions at the end of vacation, too. They didn’t anticipate the low engagement pattern because the submission of the assignment was compulsory for students. Compared to teachers’ prior expectation, our analysis revealed that 1) there were some students who didn’t complete the assignment, and 2) there were not so many students who were continuously working on the assignment: most students focused on the assignment on specific days. Plus, by comparing their performance on the exam after the vacation, we found significant differences between early engagement group and low engagement group, and high engagement group and low engagement group. It implies the importance of the engagement at the beginning of the vacation.

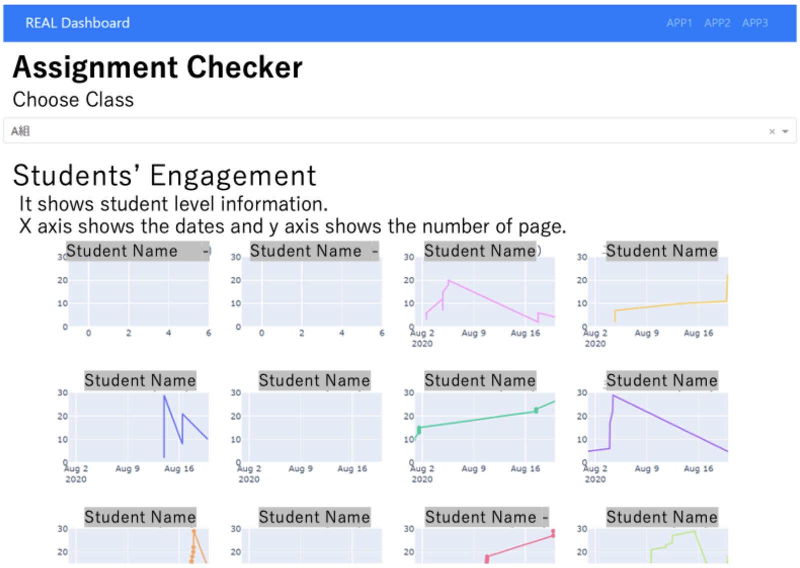
6.2 Theoretical Interpretations

It is often said that online engagement on an assignment affects students' academic performance. For instance, a study investigated predictors of students' weekly achievement and found that time spent on homework and labs were more strongly related to their performance than time spent on discussion boards or books (DeBoer & Breslow, 2014). Another research insisted that the features regarding assignment features such as average submission lead time and total quiz submission plays important roles in the dropout prediction model in MOOCs (Gardner & Brooks, 2018). However, unlike our study, the timing of the engagement was not often considered in the performance prediction before. It is not well known whether the time to finish the assignment is related to the academic performance of students.

In this paper, we make a hypothesis that early engagement reflects high self-regulation skills of students. Until now, many studies indicated that students' self-regulation skills contributed to high performance in MOOCs. For example, an online survey was conducted to identify the main cause of the high drop-out rate in MOOCs (Nawrot & Doucet, 2014). They revealed that the bad time management was the biggest reason among students. (Im & Kang, 2019) conducted path analysis between self-regulated learning and learning achievement from a large dataset from Korean Cyber University. They revealed that self-regulated learning was positively correlated to the participation and the participation was positively correlated to the learning achievement. There, the self-direction skill indirectly contributes to the learning performance. In our context, students who finished the assignments early can set a goal and deadline and make a plan to get there in time. It would be proof of the high self-regulation skills of the student, which lead to a high score of examination after the vacation.

6.3 System Development for Closing the Loop

In this study, we extracted the typical students' engagement patterns during the summer vacation period and explored the relationship between engagement patterns and their performance. However, just analyzing data is not enough for effective learning analytics implementation. As we mentioned in the introduction, effective learning analytics study should make a loop which includes some interventions to students based on the analysis results (Clow, 2012). Therefore, we plan an intervention study in the next summer vacation period at the target school. In particular, we plan to send messages to low engagement students by the first half of the vacation and encourage students to finish the assignment early. To do this, we prepared a real time monitoring tool for students' engagement in the next summer vacation period (see Figure 7). Through this dashboard, teachers can see the progress of the summer vacation assignment and send messages for each student. To check the effectiveness of the message, first we randomly select the students as an experimental group and send messages only to the half of the students at risk. Then, we will monitor if there are some behavioral changes and performance differences between the experimental and control group. It will contribute to closing loops for effective LA implementations at the target school.



REAL Dashboard

Assignment Checker

Choose Class

A組

Students’ Engagement

It shows student level information.

X axis shows the dates and y axis shows the number of page.

Figure 7. System Mock-up for Timely Intervention for the next Summer Vacation Period.

6.4 *Limitations and Conclusion*

Broadly speaking, there are two limitations in this study. First is the validity of our categorization. In this study, we used the number of answers in the first and second half of the vacation as the clustering features. These features give us advantages when we visualize the results of the clustering. However, in that context, we were not able to consider the rich time-series information for pattern clustering because they are aggregated features of daily engagement of students. Advanced clustering methods such as time-series clustering (Hung et al., 2015) may enable us to treat the rich time-series information of students’ answering data. Second is the lack of causality. In this study, we just took correlation between the extracted patterns and their performance, so we couldn’t say exactly if engagement patterns affected students’ learning performance. We planned an experimental plan above in order to check the causal relationship between them.

In this paper, we offered an example of a learning analytics platform implemented at actual junior-high Math class in Japan. In particular, we introduce a case in a summer vacation assignment. Summer vacation period is a temporal remote learning period in face-to-face classrooms, so the use of the technology is easier than normal face-to-face learning period. The analysis of log data proposed that 1) students’ progress patterns were categorized as four types based on the number of the engagement in the first and second half of the vacation, and 2) engagement in the beginning of the vacation was important for their academic performance after the vacation. Finally, we showed a real-time monitoring tool for closing the loop of learning analytics implementation at the target school. It will be used in the next summer vacation period in the target school. Our case will become the first model case of how to implement LA in secondary school in Japan and we hope the use of LA will increase in the secondary education sector as well as higher education institutes.

Acknowledgements

This study is supported by JST JPM-JAX20AA, JSPS 21J14514, SPIRITS 2020 of Kyoto University, JSPS 20K20131, JSPS 20H01722, JSPS 16H06304, NEDO JPNP18013, and NEDO JPNP20006. We thank all the teachers in the target school who kindly cooperated with this study.

References

Akçapinar, G., Hasnine, N. M., Majumdar, R., Chen, A. M.-R., Flaganan, B., & Ogata, H. (2020). Exploring Temporal Study Patterns in eBook-based Learning. *Proceedings of the 28th International Conference on Computers in Education*, 342–247.

Bae, C. L., & DeBusk-Lane, M. (2019). Middle school engagement profiles: Implications for motivation and achievement in science. *Learning and Individual Differences*, 74, 101753.

Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. *Office of Educational Technology, US Department of Education.* http://files.eric.ed.gov/fulltext/ED611199.pdf

Boroujeni, M. S., & Dillenbourg, P. (2018). Discovery and temporal analysis of latent study patterns in MOOC interaction sequences. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 206–215.

Clow, D. (2012). The learning analytics cycle: closing the loop effectively. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12*, 5.

Corrin, L., Scheffel, M., & Gašević, D. (2020). Learning Analytics: Pathways to Impact. *Australasian Journal of Educational Technology*, 36(6), 1–6.

DeBoer, J., & Breslow, L. (2014). Tracking progress: predictors of students' weekly achievement during a circuits and electronics MOOC. *Proceedings of the First ACM Conference on Learning @ Scale Conference*, 169–170.

Flanagan, B., & Ogata, H. (2018). Learning analytics platform in higher education in Japan. *Knowledge Management & E-Learning: An International Journal*, 10(4), 469–484.

Freeman, A., Becker, S. A., & Cummins, M. (2017). NMC/CoSN horizon report: 2017 K. *The New Media Consortium.* https://www.learntechlib.org/p/182003/

Gardner, J., & Brooks, C. (2018). Dropout Model Evaluation in MOOCs. *Proceedings of the AAAI Conference on Artificial Intelligence,* 32(1). https://ojs.aaai.org/index.php/AAAI/article/view/11392

Gunawardena, A. (2017). Brief survey of analytics in K12 and higher education. *International Journal on Innovations in Online Education*, 1(1). http://onlineinnovationsjournal.com/streams/analytics/07ea372322ccb0f5.html

Horn, M. B., & Staker, H. (2011). The rise of K-12 blended learning. *Innosight Institute*, 5. https://aurora-institute.org/wp-content/uploads/The-Rise-of-K-12-Blended-Learning.pdf

Hung, J.-L., Wang, M. C., Wang, S., Abdelrasoul, M., Li, Y., & He, W. (2015). Identifying at-risk students for early interventions—A time-series clustering approach. *IEEE Transactions on Emerging Topics in Computing*, 5(1), 45–55.

Im, T., & Kang, M. (2019). Structural relationships of factors which impact on learner achievement in online learning environment. *The International Review of Research in Open and Distributed Learning*, 20(1). https://doi.org/10.19173/irrodl.v20i1.4012

Kovanović, V., Mazziotti, C., & Lodge, J. (2020, March 31). *Special Section on Learning Analytics for Primary and Secondary Schools-Call for Papers*. https://learning-analytics.info/index.php/JLA/announcement/view/161

Law, N., & Liang, L. (2020). A Multilevel Framework and Method for Learning Analytics Integrated Learning Design. *Journal of Learning Analytics*, 7(3), 98–117.

Li, K. C., Lam, H. K., & Lam, S. S. Y. (2015). A Review of Learning Analytics in Educational Research. Technology in Education. *Technology-Mediated Proactive Learning*, 173–184.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., Ly, A., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Wild, A., Knight, P., Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2019). JASP: Graphical Statistical Software for Common Statistical Designs. *Journal of Statistical Software*, Articles, 88(2), 1–17.

Macarini, L. A., Cechinel, C., Santos, H. L. dos, Ochoa, X., Rodés, V., Alonso, G. E., Casas, A. P., & Díaz, P. (2019). Challenges on implementing Learning Analytics over countrywide K-12 data. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 441–445.

Nawrot, I., & Doucet, A. (2014). Building engagement for MOOC students. *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion. the 23rd International Conference, Seoul, Korea*. https://doi.org/10.1145/2567948.2580054

Ochoa, X., Lang, A. C., & Siemens, G. (2017). Multimodal learning analytics. *The Handbook of Learning Analytics*, 1, 129–141.

Sancho, M., Cañabate, A., & Sabate, F. (2015). Contextualizing learning analytics for secondary schools at micro level. *2015 International Conference on Interactive Collaborative and Blended Learning (ICBL)*, 70–75.

Schnitzler, K., Holzberger, D., & Seidel, T. (2020). All better than being disengaged: Student engagement patterns and their relations to academic self-concept and achievement. *European Journal of Psychology of Education*. https://doi.org/10.1007/s10212-020-00500-6

Low Adoption of Adaptive Learning Systems in Higher Education and How Can It Be Increased in Fully Online Courses

Rhodora ABADIA[a] & Sisi LIU[b]
[a]*UniSA Online, University of South Australia, Australia*
[b]*UniSA Online, University of South Australia, Australia*
*rhoda.abadia@unisa.edu.au

Abstract: Adaptive Learning Systems (ALS) aim to provide differentiated instructions at a personalized level of learning. While the number of students enrolled in a fully online learning environment is growing rapidly, the amount of personalization that an online facilitator can provide becomes limited, which increases the need for an ALS for more effective and efficient teaching and learning. Review of literatureindicates that though studies on ALS have been conducted for more than a decade, the adoption rate of ALS in higher education is still low. One of the main issues of the low adoption of ALS is facultysupport. Due to lack of comprehensive and systematic understanding of ALS, faculty in higher educationis reluctant to the changes brought by ALS and doubtful of the feasibility and applicability of ALS. To address this issue, the paper presents a pilot trial which includes a three-stage implementation model of ALS in a fully online learning higher education organization as a case study.

Keywords: Adaptive learning System, Online Higher Education, Misconceptions, Student Model

1. Introduction

Online learning has been designed so that students can have asynchronous learning. Online learning is also beneficial for students as they can learn at their own pace with the availability of online materials. With small number of students, online facilitators were able to devote their attentions to the students and would be able to tailor the contents for them (i.e., address their specific misconceptions, point them to the right resources and contents specific to the student's needs). As student numbers grow, they have to divide the academic team's attention by limiting the time to get to know the individual student's learning needs.

Students on average, perform two standard deviations better under one-on-one tutoring compared to standardised group instruction. Education professor, Benjamin Bloom described this effect of personalised instruction as the 2-sigma problem (Bloom, 1984). One approach to personalised instruction is adaptive teaching (or learning). Although the objective of adaptive teaching and learningis the same, the context to where it is used seems to be different. Adaptive teaching often refers to how teachers can respond, when necessary, to difference among students (Westwood, 2018). Using the theory behind adaptive teaching, the term adaptive learning systems (ALS) is often used to refer to thesystems that are technology-enabled and utilise data-driven approaches to customise instructions and personalise the learning experience (Gupta et al., 2020; Khosravi, Sadiq, &Gasevic, 2020; Mavroudi, Giannakos, & Krogstie, 2018; Newman, Bryant, Fleming, & Sark-isian, 2016; US Department of Education, 2017; Walkington, 2013). The idea of improving learning through adaptive personalisation of teaching becomes a teaching and learning movement for change in higher education (Casarez, 2019).

This paper first summarises the advantages of ALS and investigates the reasons for its current low adoption in higher education. It is reflected from the literature that faculty support issue due to lack of comprehensive and systematic understanding of ALS remains to be the major issue that leads to thelow adoption of ALS. To address this issue, the paper further proposes a three-stage implementation model in the initial phase of adopting ALS in UniSA Online as a case study for in-depth analysis.

2. Adaptive Learning Systems and Its Low Adoption in Higher Education

2.1 Adaptive Learning Systems (ALS)

ALS are defined in various forms, ranging from simple systems based on a preconceived set of rules to complex systems with self-learning algorithms (Mirata & Bergamin, 2019); or it can be a platform on which to build and contain adaptive courseware (Vignare et al, 2018). When applied to online learning, the technologies that we have now provide online facilitatorsthe ability to support each student with personalised learning and adaptability that was difficult to accomplish before. Adaptive learning may help to address the drawbacks of large online courses, a potential source of inequity (e.g., only students who frequently ask for help and post questions get attention; students at risk of failing are given more attention and thus no time left to help other students to seek their potentials). Early studies have shown that adaptive learning systems can promote student engagement (Khosravi et al., 2020; Li, Cui, Xu, Zhu, & Feng, 2018; Wang et al., 2020; Yakin & Linden, 2021); have positive effects on grade performance (Holthaus, Pancar, & Bergamin, 2019; Liu, Mckelroy, Corliss, & Carrigan, 2017; Xie, Chu, Hwang, & Wang, 2019; Yakin & Linden,2021); reduce drop-out rates (Daines, Troka, & Santiago, 2016; Mosia, 2020), and promote equity through addressing diverse students' needs and social and education background (Wang et al, 2020). There is also a growing support for leaders in higher education toward adaptive learning (Green, 2018). Despite positive attitudes of these leaders towards the adoption of adaptive learning and the growing number of research studies showing positive benefits of this approach, there is very little adoption and implementation is limited. A survey conducted by Green (2018) showed that only 8% of higher education courses use adaptive learning systems in practice and that actual adoption of innovative practices already proven to enhance undergraduate education remain low (Hariri &Roberts, 2015; Phua & Ng, 2019).

2.2 Low Adoption of Adaptive Learning Systems

Despite a massive number of studies have been conducted on ALS, there has been a notable lack of successfully implemented adaptive technology-based learning systems in practice (Cavanagh et al., 2020). This gap between research and successful application of the innovation is the so-called valley of death, which can be caused by technological and scalability-related issues, lack of resources and support from stakeholders, and pedagogical issues (Mirata & Bergamin, 2019; Samuelsen, Chen & Wasson 2019; Imhof et al, 2020).

An ALS introduces challenges to technology because it collects and analyses multiple streamsof data in real time (Zlobaite et al, 2012) and requires data integration (Samuelsen et al. ,2019) and the availability of scalable architectural and technological solutions, (Dziuban et al., 2018, Venu & Kurra, 2017). For online learning, although the Learning Management System(LMS) is capable to provide real-time data, data integration is still not fully automated. In a study conducted by Samuelsen et al (2019) on data integration, most of the research reviewed show that higher education uses multiple data sources for their learning analytics and do not integrate data but rather analyse them separately. It is crucial to have integrated data available and that integration of data is automated so that real time data can be used by the adaptive learning systems. The integration of the adaptive system in the LMS is also challenging since LMS offers pre-defined settings that requires extensive customisation (Boticario, Santos & Rosmalen, 2005; Venu & Kurra, 2017). In providing customisation, many of the algorithmsfor adaptations, results and data that can assist in the implementation of ALS remain proprietary (Johanes & Lagerstrom, 2017). In addition, since private student data is being used, how educational technologies store and/or access data must also be taken into consideration.

Other reasons for low adoption of ALS in the higher education is the lack of usability. It has been reported that the users have "counter-intuitive user experience" (Adeyemo, 2018) and usability issues for students (Imhof et al, 2020; Dziuban et al., 2017; Hariyanto, Triyono, Köhler, 2020) and teachers (Lerís et al., 2017). Although several studies have been undertaken on the usability of ALS from students' perspective, there seem to be limited studies on the usability from the teachers' perspective.

The issue with time, resources and strategic vision due to complexity, high cost and scepticism from stakeholders prevents the potential of ALS in higher education to be fully recognised. Buy-in from stakeholders is one of the major reasons of low adoption of adaptive learning systems in higher education. Teacher participation is a principal factor in adoption. Often, there is resistance to adopt by teachers (Mirata & Bergamin, 2019), teacher engagement issues and scepticism that new practices take significant effort (Tagg, 2012). The adoption of ALS also changes the role of the teachers from "telling" to designing, orchestrating and supporting learning experiences (U.S. Department of Education, 2017)and teachers are often scared to adopt this. Fears that adaptive learning is changing their role in highereducation can be traced from the idea that technology replacing instructors (Dutton, 2018).

Aside from convincing the students and teachers of the value of ALS, institutional commitment is very important. Unless mandated by an institution, 100% adoption of ALS by all teachers in higher education is impossible (Casarez, 2019). Integration of ALS into overall university strategy requires resources both financial and personnel. The institution should recognise that the implementation of ALS is a time-consuming process, and it should be factored-in when resources are allocated. Mirata and Bergamin (2019) stress the importance of securing monetary resources in addition to convincing all stakeholders. Highlighted as one of the downsides of ALS is that it is costly and requires extensive support typically beyond the expertise of the teachers. It requires more support staff in the form of instruction designers and technology specialists (Fahmy, 2004). In higher education, the concepts of cost, quality and access exist in tension with each other (King & South, 2017; Murray & Pérez, 2015) and these has always cost conflicts (not only in the adoption of ALS) but in improvements in higher education in general. "Increasing access can dilute quality. improving quality leads to an increase in costs; and reducing costs can negatively impact quality and access" (Duncan, 2009).

Another major obstacle in the adoption of ALS is the substantial development cost and effort in developing high quality content for learning. Although several ALS studies have shown its effectiveness on specific courses or context, when implemented on a bigger scale, there is lack of significant student success due to time factor and design flaws (Liu et al., 2017). Often, developed learning materials are difficult to reuse and adapt to new and different educational contexts and the difficulty of designing due to the need for the granularity of the course to be consistent.

3. Proposed Adoption of Adaptive Learning in Online Higher Education: A Case Study Approach

The practical implications of adaptive learning are currently limited since there are still various challenges that ALS are facing now. These limitations often prevent the adoption of ALS even though the current technologies have the capabilities and have multitudes of possibilities. As Casarez (2019) stated in his research that the development and success of adaptive learning as an innovation in higher education is being highly dependent on the institutional and economic environment and actors. In this paper, we proposed a staged approach in implementing adaptive teaching and learning, prioritising what is feasible in the current context of the education environment and implement automation only when all concerns and limitations identified above have been fully addressed.

Each education environment, especially online learning environment, is unique and has different factors that affect the adoption of ALS and that is the reason why context is important in this study. The case study below describes the context, motivation and introduces adaptive learning concepts in online learning.

3.1 University of South Australia Online (UniSA Online)

UniSA Online is one of the fastest growing online education providers in Australia. It is an education unit of the university that offers a wide range of undergraduate degrees designed for 100% online learning. Its online environment allows students to access the course content, weekly activities and assessments fully online and communicate with academic staff and peers through various online communication channels. The full-online learning allows flexibility for a wider range of students to continue to study. Aside from a growing demand in fully online studies, the high level of growth in its enrolment in the past few years is attributed to its high level of support provided to the students.

3.3.1 Current State

UniSA Online uses a LMS to deliver and manage contents, communication, grading, and administration of the online courses. LMS is critical in the development of adaptive learning, as they are the platforms on which most adaptive learning systems are built. In addition to LMS, a dashboard which provides real-time learning analytics is currently being used. Learning analytics collects and reports on the context of the student engagement which can be used to inform ALS. Having the learning analytics allows the online facilitators to have a new role of observing the information from the learning analytics, which suggest when and where to intervene. Although being implemented at UniSA Online, there is still an urgent need for more-fine-grained learning analytics data reporting. Learning analytics systems and tools are used to make use of insights from learning data for online facilitators to implement academic support to students. For each online course, a teaching dashboard that works based on real- time live data, captured in the normal course of a student's learning engagement, is available to the online facilitators. The reported data is analysed with the intention of optimising the student learning experience itself. The academic team often uses the dashboard to look at student engagement, follows-up with non-engaging students, and monitors student's performance. It is also used for evaluating resources and activities used in the course and looking at their impacts and effectiveness. For example, how many students and how many times the students have accessed the e-readings, videos, quizzes, and other resources and activities in the course. In addition, students are also provided the opportunity to have scheduled group drop-in sessions or an individual one-on-one online consultation. These data are then reflected in the course evaluation to inform the need to further improve the resources and evaluate the appropriateness of the learning activities implemented in the courses.

UniSA Online has been successfully providing individualised support to its students at its infancy. In the past years, as the number of online students has been growing significantly, the demand for personalised experienced also increases. Existing students expect the same personalise learning experience they had when there were smaller number of students in the course and the academic staff would like to continue providing those personalised teaching to all their students. The requests for one-on-one online consultation, or questions posted in different online communication specific to their personal learning have taken up significant resources in terms of the academic team's time, which also has brought budgetary implications and concerns for the institution.

An adaptive learning system is an appropriate solution to these concerns. The online academic staff have been initiating adaptive teaching approaches and have started to look at how ALS can be fully utilised and practiced.

3.2 Pilot Trial of the Adoption of Adaptive Learning Systems

All ALS follow a similar core architecture, the student model, domain model and the adaptive model. The student model gathers data about the student and represents the student's characteristics while the domain model refers to the content and structure of the topic to be taught. The adaptivemodel uses information from the student model and domain model to provide a set of recommend learning activities and tailored feedback. This can be simplified by looking at the design process: gather data from the student, model the student, where and when can adaptation be applied and how will adaptation be provided. In this case study, the same design process is followed. The reason behind the staged implementation of the processes is that the students from the courses can start getting the benefit of adaptive learning even though it has not been fully automated and implemented.The proposed implementation of each process and how it addresses the issues of faculty support will be described.

A proof of concept is needed to get buy-in for stakeholders and in few cases where adoption of adaptive learning systems in higher education was successful, participation of the faculty is an important factor (Dziuban et al., 2018). This proof of concept can be staged and ease budgetary requirements.

A key part on the adoption of ALS depended on the support of the faculty therefore they shouldbe involved from the start of the process. Several issues in the previous sections were identified on why there were faculty scepticisms. In the proposed stages, the aim is not only to build a proof of concept but also alleviate these scepticisms.

In this pilot trial, the online facilitators from various disciplines, psychology, design, accounting, and programming, were involved in identifying how adaptive learning can assist in improving student engagement, motivation, retention and performance in their courses. We recognised that successful adoption of innovation in an organisation depends on its unit's culture and is influenced by its employees (Hogan & Coote, 2014), and would be adopting this concept in the phased implementation of ALS.The focus is to simulate the adaptive learning systems process and automate the lesser resource intensive part of adaptive systems (i.e., estimation of students' progress) since data from learning analytics and LMS are already available. The online facilitators are involved in identifying what student data can be pulled and used in its raw form from the LMS, which data is available from LMS but still needs data formatting and processing. In this stage, the estimation of student's progress is granular and looks at the data that is an indicator of student's learning, i.e., misconceptions, marking criteria, feedback and marks on summative assessments. In this study, misconceptions are defined as inaccurate or incomplete ideas about a concept or a process (Savion, 2009). Given the data about the students' summative assessments, a clustering-based approach is applied to estimate the students' progress based on misconceptions. From these groups of misconceptions, the online facilitators identify the appropriate intervention. The online facilitator's participation in this process will help them understand the changing role but also remove the fear that technology will be replacing them and the fear of the use of the technology. They will also be aware of the amount of effort needed in developing ALS, understand why automation is necessary and more efficient, and experience the effectiveness of the approach. Figure 1 shows the three stages. These stages are iterative such that continuous refinements of the process can be implemented after the initial phase

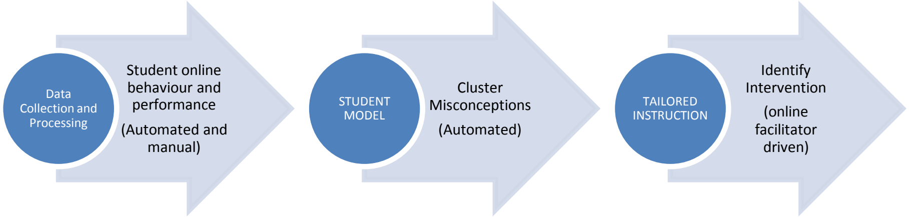


Figure 1. Proposed three-staged Approach in the Initial Phase of Adoption of ALS at Unisa Online.

3.2.1 Data Collection and Processing

The LMS that interacts with students to deliver content and assessments to support student learning captures time-stamped student input and behaviours within the system. Since at the initial phase, the student modelling will only focus on identifying student's misconceptions, the data will focus on the student's summative assessments grades and feedback from online facilitators. The online facilitators from different disciplines, psychology, digital media, accounting, and programming will be involved in the collection of data from their courses. These data are available in the LMS but are in differentforms. For example, grades can be downloaded as CSV file while the marking criteria and feedback for the assessment are in the document form. Students grades in the assessments are numerical and can be extracted from the grading book. The courses from different disciplines that is used in this initial phase are introductory courses. For programming courses, especially for the introductory ones, syntax, logic and semantic errors remain to be the main barriers for students to fully grasp the language-independent concepts and learn how to code novel problems independently (Sanati, Soon, & Lin, 2020; Veerasamy, D'Souza, &Laakso, 2016).

In this study, the data collected are the occurrences of common logical, syntactical and semantics errors identified as part of marking criteria and feedback for summative assessments. For each type of misconception, errors are further grouped into three levels of severity and saved in a spreadsheet. For non-programming courses, the marking criteria and feedback are in the form of documents. It is tedious to manually go through all the qualitative comments and extract the feedback. Natural Language Processing (NLP) is used to manipulate the textual information for semantics and syntactic analysis. The result of this is a text data set that contain the summary contents and extracted key phrases relevant to the assessment. The result of the data collection and processing also includes a set misconception patterns that have been extracted from the text analysis.

3.2.2 Student Model

In this stage, the student model focuses in estimating the student's progress using clustering algorithm based on student misconceptions. We use the term misconception when a student's knowledge is erroneous, illogical or misinformed. In some adaptive learning systems, the student's demographics and learning behaviour data are used in tracking the student's progress. An initial research on students' demographics at UniSA Online (Bretana et al., 2020) indicates that these data are not significant factors in predicting a student's success in the course, but it is the performance-related data that yield a more accurate prediction. This is the reason why at the initial phase of this study, the focus will be on the performance, specifically misconceptions in summative assessments.

The rationale behind clustering students based on misconceptions rather than their achievement (or grades) is that research in education have shown that grouping students based only on their achievement or ability (grades) is not an effective strategy for improving educational outcomes (Francome & Hewitt, 2020; Steenbergen-H, Make & Olszewski-Kubilius, 2016). In an "ideal ALS" scenario, an individual student model is identified and personalisation of the correction of the misconception is applied. However, in this initial phase, since intervention is not fully automated but online facilitator driven, grouping students with similar misconceptions is a more efficient way to use the resources (time, money and academic staff). By gaining insight on students' misconceptions, these misconceptions can be addressed irrespective of the grades the student's grade. This gives all students with different abilities to achieve better.

The clustering-based approach uses clustering algorithms to group students based on their misconceptions. For example, in the programming course, the data collected containing the syntactical, semantical and logical errors in the summative assessment will be used as data points. The clustering algorithm will then generate student groups using similarity measurements. For other courses, the common misconceptions found in the data processing stage, will also be used as data points and compute for similarity measurement to form student groups with similar misconceptions. Clustering isan iterative process. After several iterations, the final clusters are obtained when the error (sum of square errors) remains unchanged. The common patterns of group of misconceptions can be deduced from observing the key features for each cluster. This will assist the faculty in preparing for intervention.

3.2.3 Tailored Instruction

Due to the substantial cost and effort in developing adaptive content, in the initial phase, the intervention is driven by the online facilitators. Once the students are clustered based on their misconceptions, the online facilitators will identify the appropriate intervention for the misconceptions. With small number of students, prior to the adoption of the proposed approach in this paper, the online facilitators will schedule a one-on-one student consultation with students to address the misconception(s). This approach is not scalable and resource intensive. As an initial intervention approach, the online facilitators will conduct a "targeted online session". Students within the same cluster are invited to attend the session where the online facilitatorsfocus on the specific needs and interests of the group of students and take them incrementally to the next level of learning. This approach is consistent with the concept of tutoring which exemplifies the essence of effective teaching (Coffey, 2016).

An example of how targeted online sessions are conducted includes the online facilitators engaging the group of students in activity-based methods where they are shown conflicting events or examples, which challenge each other's answers or the student's own misconception(s). This is consistent with Longfield's (2009) study showing that the approach described above results in more lasting learning. Another approach used in the online targeted sessions are carefully selected demonstrations to help students identify the causes of their misconceptions and correct them in a moreeffective manner. In addition to this, common misconceptions are revisited to help students reconstruct their conceptual framework.

3.3 Evaluation

Formative evaluation takes place during the implementation of the stages so that corrective action can be done as problems arise. Formative reports will be collected in each of the stages in the pilot project. Summative evaluation is performed once the project has been completed. Summative evaluation of the pilot trial will use the Context, Inputs, Processes, and Products (CIPP) Evaluation Model (Stufflebeam& Zhang, 2017). Figure 2 shows how CIPP Evaluation Model is used in the pilot project. When applied to this project, this model systematically collects information about the project to identify strengths and limitation of the process, to improve the project's effectiveness and plan for the next phase. Both formative and summative evaluations will be performed.

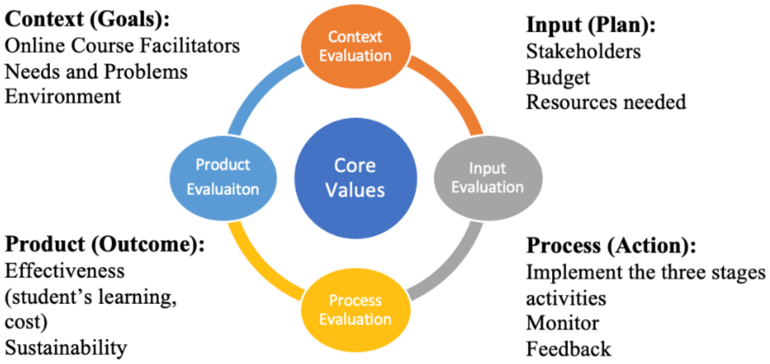


Figure 2. Implementation of Context, Input, Process and Product Evaluation Model.

In the context component, the goal is to define the relevant context, assess the needs of online facilitators underlying the needs and check whether the aim of this project responds to the assessed needs. A survey will be given to online facilitators identifying their views in personalised learning; how are they implementing personalised academic support; and what are the barriers for them in implementing personalised teaching. From the result, a comparison on how the project is responding to their needs. A needs-gap analysis (Upadhyaya, 2013) is performed to better fit the goal to the current needs and re-align the process with the strategy to meet the goal.

In the input component, the pilot trial involving the three stages is assessed. We look at resources needed, budget and cost to meet the needs and achieve the goal of this project. Process evaluation assesses the implementation of the plans. It determines whether the activities in all the proposed stages in the pilot trial have been implemented as intended and resulted inexpected outputs. The formative evaluation conducted throughout the implementation of the pilot project will assist in the review of the activities and expected outputs of the processes. Feedback is done throughout the implementation of the plan and then checks the extent of the plan that was completed. The fourth component is product evaluation. In education, an effective evaluation of new teaching practice is by looking at the student outcomes. The effectiveness of the three-staged approachin student's learning is evaluated by looking at the student's performance in the proceeding related summative assessments. Some examples include looking at student's performance in the final exam on the specific topic where intervention was applied; or at assessments that require the synthesis of skills and knowledge which is preceded by the summative assessment were the misconceptions wereidentified. In addition, the project is assessed based on impact, cost-effectiveness and sustainability of the approach.

The result of the evaluation will then provide a comprehensive review of the value of the project by having a report on the quality, positive and negative outcomes and impacts and cost- effectiveness.

3.4 Preliminary Findings

We have applied the three-stage approach to a fully-online introductory programming course for preliminary results and findings. Data was collected from the marking feedback of 163 students who submitted the first programming assessment. Coding errors that each student made were identified as either syntactical, logical and semantics errors. We fed the transformed data into the clustering-based approach (discussed in section 3.2.2) and obtained three clusters with the following details: the first cluster contained 35 students who performed below average in syntax area and poor in the semantic and very poor in logic; the second cluster contained 80 students who mainly performed average in syntax and logic areas but performed poorly in semantic area; and the third cluster contained 48 students who performed well above average in all three areas. Interventions were implemented based on the first two clusters. A revision online session was offered to the first cluster, which is an online tutorial on control structures, and discussion of logic and semantic errors found in the assessment. The second cluster group have been given additional resources on semantical errors and a few debugging exercises related to semantics. Since this is not a required activity, there were very few participants in the intervention activities. Comparing the performance of students in the first assessment against the final programming assessment developed on the same learning objectives, four of the nine in the first cluster had improved results, from F1 to HD, F2 to C, P1 to P2, D to HD, respectively. One of the two in the second cluster who attempted the additional exercises has shown improved result in the final assessment (i.e, D to HD)

Though the above findings indicate that to some extent, the proposed three-stage approach assists the online facilitators in addressing students' misconceptions through a clustering-based approach and providing tailored instructions to improve student performance, a solid conclusion is yet to be drawn due to the overall low participants in the comparative analysis.

4. Conclusion

This paper explored the current state of ALS in higher education and the need for its integration in fully online courses. Review of literature indicates that there are several issues, such as obstacles to data acquisition and availability, lack of usability, complexity, high cost and scepticism from stakeholders, that lead to the low adoption of ALS in the higher education industry. One of the principal reasons for low adoption addressed in this paper is buy-in from the faculty. We proposed a three-staged approach to the implementation of ALS using a case study and discussed how online facilitators can be involved to have faculty participation at the start and have increased support in the adoption of ALS. The three stages are data collection, student model and tailored instruction. The implementation of the proposed approach is evaluated using the CIPP evaluation model. The result of this evaluation identifies the value of the project and helps plan for the next phase.

In this pilot phase, data collection and processing are a combination of manual and automated process, student modelling used clustering and tailored instruction is driven by the online course facilitators. To prepare the dataset for generating student clusters, we converted the common misconceptions observed in the summative assessments into three types of errors, including syntax, logic and semantic. A clustering-based approach was utilised to group students based on their performance determined by the occurrence of errors and severity level. According to the error patternsderived from the clusters, online course facilitators then developed tailored instruction and provided each student group with more targeted intervention and support.

The next step for this study is to look at the evaluation results and investigate the implementation of fully automated data collection and processing. This requires looking at the both the usability and administrative requirements for online course facilitators and the technical feasibility of incorporating this in the current learning environment. The evaluation results will also report on the effectiveness of the clustering approach and identify if this is the best alternative option while adaptive content is still not fully automated, or if there is a need to individualise student modelling if the clustering approach is already effective. Given the data from the implementation of the pilot study, theinitial identification of learning materials for reuse and adaptation will be explored. The feedback provided to students in the targeted sessions will be analysed and translated to contents that can be used in automating and tailoring instructions. Notably, this study is part of a larger research that aims at the full adoption of ALS in the organisation described in the case study. As part of the proof of concept being developed to help the business case, this study is undertaken to secure the resources and institutional commitment.

References

Adeyemo, A. (2018, October 7). *Reduced-seating format frustrates UCF College of Business students*. Retrieved September 24, 2019, from NSM.today website: http://www.nicholsonstudentmedia.com/life/reduced-seating-format-frustrates-ucfcollege-of-business-students/article_b2a83090-c04d-11e8-a9b0-ab8c800dc2db.html

Bloom, B. S. (1984). The 2-sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher, 13* (6), 4–16.

Boticario, J., Santos, O., & Rosmalen, P. (2005). Technological and management issues in providing adaptive education in distance learning universities. *EADTU. Retrieved Dec 19,* 2007.

Bretana, N. A., Robati, M., Rawat, A., Pandey, A., Khatri, S., Kaushal, K., . . . Abadia, R. (n.d.). Predicting student success for programming courses in a fully online learning environment.

Casarez, R. R. (2019). *Adaptive learning: Dilemmas of automated instruction in postsecondary education*. Michigan State University.

Cavanagh, T., Chen, B., Lahcen, R. A. M., & Paradiso, J. R. (2020). Construct- ing a design framework and pedagogical approach for adaptive learning in higher education: A practitioner's perspective. *International Review of Research in Open and Distributed Learning, 21* (1), 172–196.

Coffey, D. (2016), Tutoring, *The SAGE Encyclopedia of Online Education*, pp. 1139–1144.

Daines, J., Troka, T., & Santiago, J. (2016). Improving performance in trigonometry and pre-calculus by incorporating adaptive learning tech- nology into blended models on campus. In *123rd annual asee conference & exposition, new orleans, louisiana* (Vol. 10, p. 25624).

Duncan, A. (2009). Rethinking Higher Education: Moving beyond the Iron Triangle. Trusteeship, 17(5), 8–11.

Dutton, G. (2018). Adapting to Adaptive Learning. Training Magazine. Retrieved from https://trainingmag.com/trgmag-article/adapting-adaptive-learning/

Dziuban, C., Moskal, P., Johnson, C., & Evans, D. (2017). Adaptive learning: A tale of two contexts. *Current Issues in Emerging eLearning, 4* (1), 3.

Dziuban, C., Moskal, P., Parker, L., Campbell, M., Howlin, C., & Johnson, C. (2018). Adaptive learning: A stabilizing influence across disciplines and universities. *Online Learning, 22* (3), 7–39.

Fahmy, M. F. (2004). Thinking about technology effects on higher education. *Journal of Technology Studies, 30* (1), 53–58.

Francome, T., & Hewitt, D. (2020). "My math lessons are all about learning from your mistakes": how mixed-attainment mathematics grouping affects the way students experience mathematics. *Educational Review, 72* (4), 475–494.

Green, K. (2018). *Campus computing 2018: The 29th national survey of computing and information technologyin American higher education*. Retrieved from https://www.campuscomputing.net/content/2018/10/31/the-2018-campus-computing-survey

Gupta, S., Ojeh, N., Sa, B., Majumder, M. A. A., Singh, K., & Adams, O. P. (2020). Use of an adaptive e-learning platform as a formative assessment tool in the cardiovascular system course component of an mbbs pro- gramme. *Advances in Medical Education and Practice, 11,* 989.

Hariri, A., Roberts, P., et al. (2015). Adoption of innovation within universities: Proposing and testing an initial model. *Creative Education, 6* (02), 186.

Hariyanto, D., Triyono, M. B., & K¨ohler, T. (2020). Usability evaluation of per- sonalized adaptive e-learning system using use questionnaire. *Knowledge Management & E-Learning: An International Journal, 12* (1), 85–105.

Hogan, S. J., & Coote, L. V. (2014). Organizational culture, innovation, and performance: A test of schein's model. *Journal of business research, 67* (8), 1609–1621.

Holthaus, M., Pancar, T., & Bergamin, P. (2019). Recommendation acceptance in a simple adaptive learning system.

Imhof, C., Bergamin, P., & McGarrity, S. (2020). Implementation of adaptive learning systems: Current state and potential. In *Online teaching and learning in higher education* (pp. 93–115). Springer.

Johanes, P., & Lagerstrom, L. (2017). Adaptive learning: The premise, promise, and pitfalls. In *Proceedings of the 124th asee annual conference and expo- sition.*

Khosravi, H., Sadiq, S., & Gasevic, D. (2020). Development and adoption of an adaptive learning system: reflections and lessons learned. In *Proceedings of the 51st acm technical symposium on computer science education* (pp. 58–64).

King, J., & South, J. (2017). Reimagining the role of technology in higher education: A supplement to the national education technology plan. *US Department of Education, Office of Educational Technology.*

Ler´ıs, D., Sein-Echaluce, M. L., Hern´andez, M., & Bueno, C. (2017). Validation of indicators for implementing an adaptive platform for moocs. *Computers in Human Behavior, 72,* 783–795.

Li, H., Cui, W., Xu, Z., Zhu, Z., & Feng, M. (2018). Yixue adaptive learning system and its promise on improving student learning. In *Csedu (2)* (pp. 45–52).

Liu, M., McKelroy, E., Corliss, S. B., & Carrigan, J. (2017). Investigating the effect of an adaptive learning intervention on students' learning. *Educa- tional technology research and development, 65* (6), 1605–1625.

Longfield, J. (2009). Discrepant teaching events: Using an inquiry stance to address students' misconceptions. *International Journal of Teaching and Learning in Higher Education, 21*(2), 266-271.

Mavroudi, A., Giannakos, M., & Krogstie, J. (2018). Supporting adaptive learning pathways through the use of learning analytics: developments, challenges and future opportunities. *Interactive Learning Environments, 26* (2), 206–220.

Mirata, V., & Bergamin, P. (2019). Developing an implementation framework for adaptive learning: A case study approach.

Mosia, N. (2020). Adaptive learning implementation–a cognitive description experiment for first year engineering students at a distance education uni- versity. In *Eden conference proceedings* (pp. 124–130).

Murray MC, Pérez J (2015). Informing and performing: a study comparing adaptive learning to traditional learning. Inf Sci 18:111Newman, A., Bryant, G., Fleming, B., & Sarkisian, L. (2016). Learning to adapt 2.0: the evolution of adaptive learning in higher education. *Tyton Partners white paper*.

Phua, J., Yeo, E., & Ng, S. (2019). Understanding teaching and learning practices of online adaptive mathematics tutoring platform. In *Proceedings of the ninth international conference on learning analytics& knowledge (lak'19). society for learning analytics research.*

Samuelsen, J., Chen, W., & Wasson, B. (2019). Integrating multiple data sources for learning analytics—reviewof literature. *Research and Practice in Technology Enhanced Learning, 14* (1), 1–20.

Sanati, F., Soon, L., & Lin, Y. (2020). Intelligent teaching and learning plat- form for introductory programming subjects. In *Proceedings of the 12th international conference on computer modeling and simulation* (pp. 3–8).

Savion, L. (2009). Clinging to discredited beliefs: The larger cognitive story. *Journal of the Scholarship ofTeaching and Learning, 9,* 81-92.

Steenbergen-Hu, S., Makel, M. C., & Olszewski-Kubilius, P. (2016). What one hundred years of research says about the effects of ability grouping and acceleration on k–12 students' academic achievement: Findings of two second-order meta-analyses. *Review of Educational Research, 86* (4), 849–899.

Stufflebeam, D. L., & Zhang, G. (2017). *The cipp evaluation model: How to evaluate for improvement and accountability*. Guilford Publications.

Sunkara, V. M., & Kurra, R. R. (2017). A learner-centric personalized and adaptive e-learning framework for higher education. *International Journal of Advanced Research in Computer Science, 8* (5).

Tagg, J. (2012). Why does the faculty resist change? *Change: The Magazine of Higher Learning, 44* (1), 6–15.

Upadhyaya, Makarand. (2013). Customer Satisfaction Measurement: an empirical Study of the Need – Gap Analysis in the Service Industry. *Journal of Economics and Business Research, 19*(2), 54–61.

US Department of Education. (2017). Reimagining the role of technology in ed- ucation: 2017 national education technology plan update. *US Department of Education*.

Veerasamy, A. K., D'Souza, D., & Laakso, M.-J. (2016). Identifying novice student programming misconceptions and errors from summative assess- ments. *Journal of Educational Technology Systems, 45*(1), 50–73.

Vignare, K., Lammers Cole, E., Greenwood, J., Buchan, T., Tesene, M., De- Gruyter, J., . . . others (2018). *A guide for implementing adaptive course- ware: From planning through scaling.* Retrieved from Joint publication of Association of Public and Land-grant . . ..

Walkington, C. A. (2013). Using adaptive learning technologies to personal- ize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology, 105* (4), 932.

Wang, S., Christensen, C., Cui, W., Tong, R., Yarnall, L., Shear, L., & Feng, M. (2020). When adaptive learningis effective learning: comparison of an adaptive learning system to teacher-led instruction. *Interactive Learning Environments*, 1–11.

Westwood, P. (2018). *Inclusive and adaptive teaching: Meeting the challenge of diversity in the classroom*. Routledge.

Xie, H., Chu, H.-C., Hwang, G.-J., & Wang, C.-C. (2019). Trends and de- velopment in technology-enhanced adaptive/personalized learning: A sys- tematic review of journal publications from 2007 to 2017. *Computers & Education, 140*, 103599.

Yakin, M., & Linden, K. (2021). Adaptive e-learning platforms can improve student performance and engagement in dental education. *Journal of Dental Education*.

Zliobaite, I., Bifet, A., Gaber, M., Gabrys, B., Gama, J., Minku, L., & Musial,K. (2012). Next challenges for adaptive learning systems. *ACM SIGKDD Explorations Newsletter, 14* (1), 48–55.

GUI-Based System for Effortless Program-Visualization Creation Using Time-Series Information

Koichi YAMASHITA[a*], Miyu SUZUKI[b], Satoru KOGURE[b], Yasuhiro NOGUCHI[b], Raiya YAMAMOTO[c], Tatsuhiro KONISHI[b] & Yukihiro ITOH[d]

[a]*Faculty of Business Administration, Tokoha University, Japan*
[b]*Faculty of Informatics, Shizuoka University, Japan*
[c]*Faculty of Engineering, Sanyo-Onoda City University, Japan*
[d]*Shizuoka University, Japan*
*yamasita@hm.tokoha-u.ac.jp

Abstract: In this paper, we describe a system that effortlessly creates program visualization (PV) by incorporating time-series information into a graphical user interface (GUI) system for PV creation. Although several PV systems have been developed, only a few have been introduced or used continuously in actual classes. One of the main obstacles to using PV systems in actual classrooms is the significant amount of time needed to integrate PV systems into actual educational setups. We developed a PV system called TEDViT and introduced it into several practical classes. While programming learning with TEDViT had a noticeable effect, the time required for PV customization was a non-trivial problem. To address this issue, GUI-based WYSIWYG PV editor would be a promising approach. However, many existing systems only support PV drawing. We believe that PVs should be more than mere drawings of data structures. They should be sequences of drawings with a program-execution process. This study has therefore developed a PV-creation support system that considers the continuity of drawings by incorporating time-series information into the GUI. An evaluation experiment was conducted to measure the time required to create PVs using our system. The results suggest that our GUI system noticeably improves the efficiency of PV creation.

Keywords: Programming education, program visualization system, program visualization design, educational authoring tool

1. Introduction

Program visualization (PV) is a widely accepted approach for supporting novice learners who find it difficult to obtain a clear image of program behavior. Thus far, several PV systems have been developed, and many positive learning effects have been reported (Pears et al., 2007). However, few PV systems have been introduced continuously in actual classes. One of the main obstacles to continuous use is the time cost involved; teachers who introduce PV systems into their classrooms must design, integrate, and maintain the PV generated by the system alongside their own lesson plans.

To address this issue, we have developed a PV system, the Teacher's Explaining Design Visualization Tool (TEDViT) and conducted several classroom practice sessions using this system (Kogure et al., 2014). One distinctive feature of TEDViT is the fact that it enables teachers to customize PVs, based on their own instruction plans. Through this feature, teachers can design, integrate, and maintain PVs that reflect their own intentions, achieving positive evaluation results that suggest significant learning effects. Compared to existing PV systems, however, this feature incurs additional costs for PV customization. More work is needed to reduce the cost of using PVs.

While many factors may account for the high cost of PV creation, this paper focuses on PV customization—the most direct way to increase the learning effectiveness of PV systems. Tezuka et al. (2016) proposed a way to reduce the cost of PV creation. They developed a GUI-based system that specified the positions and attributes of drawn objects in TEDViT, making PV creations more intuitive—and not based on the numerical specification of coordinates, as previously required. In evaluation experiments, their system reduced the time needed for PV creation by approximately 40%. It is thus a promising cost-reduction approach. However, their system only supports PV drawing. We believe that PV needs to include a sequence of drawings that follow the program execution process, rather than drawing alone.

The present study has developed a system that supports PV creation by incorporating sequences of drawings into the GUI, based on time-series information. To evaluate the effectiveness of this system, we conducted an experiment in which the subjects were asked to create PVs. This paper describes our GUI-based system for effortless PV creation and evaluation experiment. The evaluation results suggest that our approach to PV creation, based on sequences of drawings, is an effective way to support PV creation.

2. Related Works

2.1 Existing PV Systems

During the past few decades, several PV systems have been developed for novice learners, including Python Tutor (Guo, 2013), Jype (Helminen & Malmi, 2010), and PROVIT (Yan, Nakano, Hara, Suga, & HE, 2014). These systems differ in certain ways. For example, Python Tutor runs on a web browser and does not require local installation. Jype provides a learning environment that integrates the PV and automatic-assessment systems for exercise assignments. PROVIT uses 3D graphics in its visualizations. However, all of these systems are similar in the sense that they all visualize the target program and its data structures in a uniform way. Generally speaking, these systems are capable of visualizing programs using a fixed visualization policy. They also allow learners to observe changes in data structure during the execution of each statement. This function is provided by a graphical user interface (GUI), such as next/previous buttons for stepwise execution of the target program. Sorva, Karavirta, and Malmi (2013) provide a comprehensive overview of more than 40 PV systems, which share many similarities.

PV systems demonstrate the runtime behavior of computer programs to novice learners by providing PV that visually encodes data and shows how it is processed in a running program. Novice learners often find it difficult to trace program states and behaviors via data structures. By bridging the gap between their reasoning and computational processes, PV systems can improve novice learners' understanding of programs (Tudoreanu, 2003). However, as Sirkiä and Sorva (2015) have pointed out, PV systems are not always effective. Learners may struggle to understand the meaning of visual elements or neglect important aspects and focus on peripheral elements. We would argue that this reflects a failure to integrate with other materials or offer customizable systems. It also suggests a poor fit with teachers' personal pedagogical styles. Sorva, Karavirta, and Malmi (2013) call this the "problem of dissemination." Teachers' in-class explanations shape the reasoning of learners. Similarly, the designs of PV-system developers shape the way that computational processes are visualized. For PV systems to adequately bridge the gap, they must be designed, integrated, and maintained by teachers.

A few systems, such as ANIMAL (Rößling & Freisleben, 2002), can customize PVs. Prior knowledge and preparation are needed to customize PVs. ANIMAL uses the script language, AnimalScript, to define PV—and the cost of learning it is significant. Moreover, PV creation requires a non-trivial quantity of script code. The sample script for a bubble sort algorithm bundled in ANIMAL consists of 170 lines of script code. More efforts are therefore needed to reduce the cost of PV creation.

2.2 Effortless PV Creation

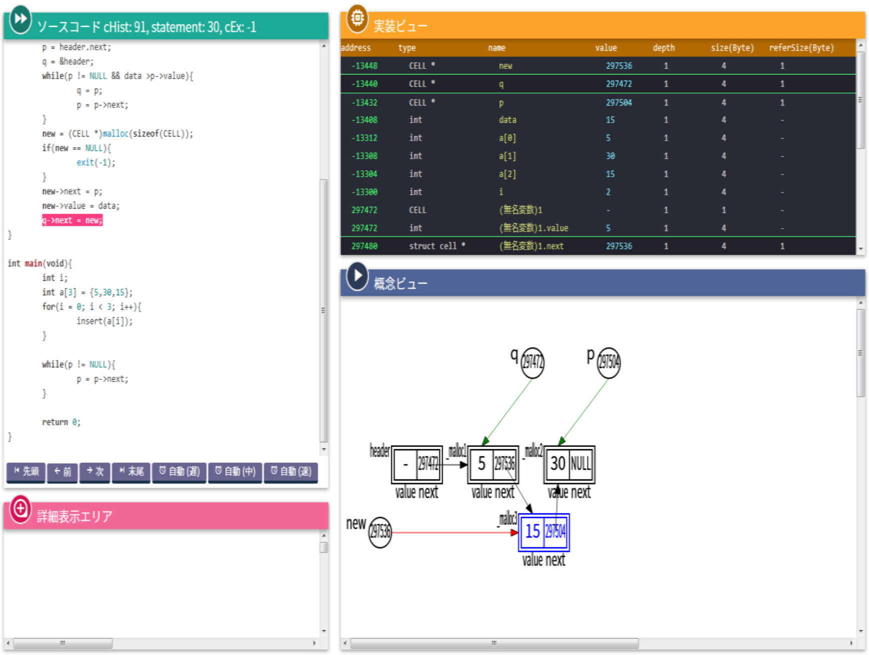
Several studies have investigated ways to reduce the cost of algorithm visualization (AV) and PV creation. Malone, Atkinson, Kosa, and Hadlock (2009) developed a pseudo-code system in which a definition of the visualization can be included in the pseudo code used to represent the target algorithm. The pseudo-code interpreter automatically derives AV from algorithm implementations. This study argues that it is essential to increase AV effectiveness and effortlessness. Velázquez-Iturbide, Pareja-Flores, and Urquiza-Fuentes (2008) developed a system that allows teachers to select PVs from an automatically generated PV sequence in a list format. Although they argue that their system improves effortlessness in PV creation, they do not present any experimental results to prove the point objectively. Rößling and Ackermann (2007) have developed a framework that allows teachers to create AV content on-the-fly by adjusting variable values and attributes. Their framework derives from a set of prepared templates that define the visualization details for various algorithms. These definitions can be saved freely, reducing the effort needed to reuse AVs. PV systems, such as Jeliot 3 (Moreno, Myller, Sutinen, & Ben-Ari, 2004) are often considered effortless because they automatically generate PVs by providing target programs only, although the PVs generated in this way cannot be customized. There are various approaches to effortless PV creation and simple comparisons are difficult to make.

Ihantola, Karavirta, Korhonen, and Nikander (2005) have defined a taxonomy to characterize effortlessness in AV systems. Based on a survey conducted among CS educators, they identify three main categories—*scope, integrability,* and *interaction*—and evaluate several existing systems. The *scope* refers to the range of contexts in which the AV system can be applied—the various algorithmic domains for which the system can be adapted. *Integrability* refers to third-party effortlessness: how easy it is to integrate the AV system into educational setups. *Interaction* is the extent to which the system can be used for different cases. This factor is based not only on interactions between AV content and learners but also on interactions between teachers and content and the extent to which the content is customizable. Although PV systems tend to provide visualizations at a lower level of abstraction than AV systems (Sorva et al., 2013), these three factors also apply to PV creation. PV systems aim to help users understand underlying algorithms by visualizing program behavior.

The present study focuses on the interactions between teachers and content, among other aspects of effortlessness in PV creation, based mainly on using TEDViT in the classroom. TEDViT is a PV system, which allows teachers to customize PV through their own instructions. The practice classes obtained positive learning effects from the interactions between teachers and content (i.e., PV customizations); we can observe students learning to understand programs (Yamashita et al., 2017), extend their class learning style (Yamashita et al., 2016), and so on. The goal of this work is to improve the effortlessness of PV creation by developing a system that supports teachers' PV customizations.

2.3 TEDViT

The TEDViT system interprets each visualization policy by scanning the configuration file and visualizing the PV accordingly. Figure 1 presents a screenshot of the learning environment visualized by TEDViT. The configuration file comprises a set of drawing rules, each of which is a comma-separated value (CSV) entry, consisting of a condition and an object. The condition defines the prerequisites needed to fire the drawing rule. Teachers can use a conditional equation (consisting of a statement ID, variables in the target program, constant values, and comparison operators) to determine the drawing timing. Here, the statement ID is a unique identifier automatically assigned to all statements in the target program by TEDViT. The object defines the operation ("create," "delete," or "update") used to edit the target object and the attributes need to draw it, which include object type, position, color, and corresponding variables. These features make PV customizable in TEDViT by allowing teachers to define the drawing rules. TEDViT supports visualizations of only C programs.



ソースコード cHist: 91, statement: 30, cEx: -1

```
        p = header.next;
        q = &header;
        while(p != NULL && data >p->value){
                q = p;
                p = p->next;
        }
        new = (CELL *)malloc(sizeof(CELL));
        if(new == NULL){
                exit(-1);
        }
        new->next = p;
        new->value = data;
        q->next = new;
}

int main(void){
        int i;
        int a[3] = {5,30,15};
        for(i = 0; i < 3; i++){
                insert(a[i]);
        }

        while(p != NULL){
                p = p->next;
        }

        return 0;
}
```

実装ビュー

address	type	name	value	depth	size(Byte)	referSize(Byte)
-13448	CELL *	new	297536	1	4	1
-13440	CELL *	q	297472	1	4	1
-13432	CELL *	p	297504	1	4	1
-13408	int	data	15	1	4	-
-13312	int	a[0]	5	1	4	-
-13308	int	a[1]	30	1	4	-
-13304	int	a[2]	15	1	4	-
-13300	int	i	2	1	4	-
297472	CELL	(無名変数)1	-	1	1	-
297472	int	(無名変数)1.value	5	1	4	-
297480	struct cell *	(無名変数)1.next	297536	1	4	1

Figure 1. Screenshot of a Learning Environment generated by TEDViT.

For example, the drawing rule shown in Figure 2 means that when the statement with ID "10" in the target program is executed, TEDViT draws a circle object and assigns it the object ID "OBJ1." The corresponding variable is "*i*"; hence, the value of *i* is drawn inside OBJ1. OBJ1 is placed in position (x1, y1) with black, white, and black as the line, background, and inner-character colors, respectively, in accordance with the values indicated in the rule. TEDViT provides buttons for stepwise control of the target-program execution, similar to the GUI in typical PV systems. When a learner clicks on the "previous" and "next" buttons, TEDViT finds the corresponding program-execution status, fires the rule for which the condition is satisfied, and visualizes the corresponding drawn objects.

```
state==10, create, OBJ1, circle, i, x1, y1, black, white, black
```

Figure 2. Example of a TEDViT drawing rule.

In classroom practice with TEDViT, the use of PVs that reflected the teacher's instructional intent led to certain positive learning effects. However, the customizability of PV in TEDViT also creates a burden for teachers, who must define the drawing rules. According to Yamashita et al. (2016), it took approximately 30 minutes to define the drawing rules for a single sample selection-sort code, consisting of approximately 30 statements. Although actual teachers rated this as an acceptable class-preparation cost, we consider it a significant burden.

2.4 GUI System for PV Creation

To support the interaction between teachers and content, some systems have functions that design PVs via GUI. Based on WYSIWYG AV editor implementation, Karavirta, Korhonen, Nikander, and Tenhunen (2002) evaluated the effortlessness of existing AV systems. Using TEDViT, Tezuka et al. (2016) developed a GUI system that visually defined the positions and attributes of drawn objects in order to reduce the cost of defining drawing rules. Hereafter, this paper will refer to their system as Tezuka GUI; Figure 3 presents a screenshot of this system.

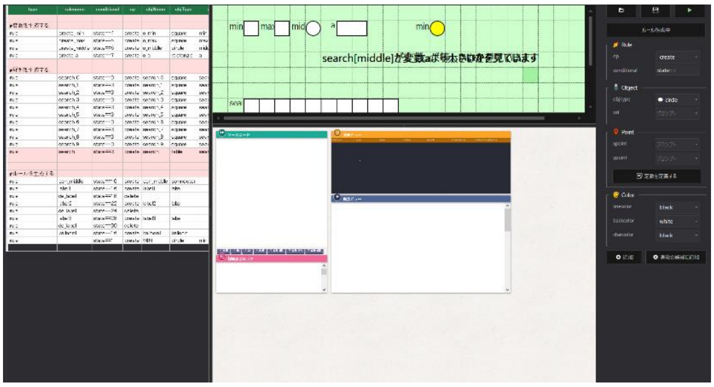


Figure 3. Screenshot of Tezuka GUI.

Tezuka GUI has functions that visually specify the positions of drawn objects, list their available attributes (line color, background color, character color, etc.), specify the values in combo-box style, and highlight grammatical errors in the drawing rules. Tezuka et al. (2016) evaluated the extent to which Tezuka GUI improved effortlessness (measured using the time needed to create drawing rules for TEDViT) and found that the measured times for rule creation were approximately 40% less with Tezuka GUI than without.

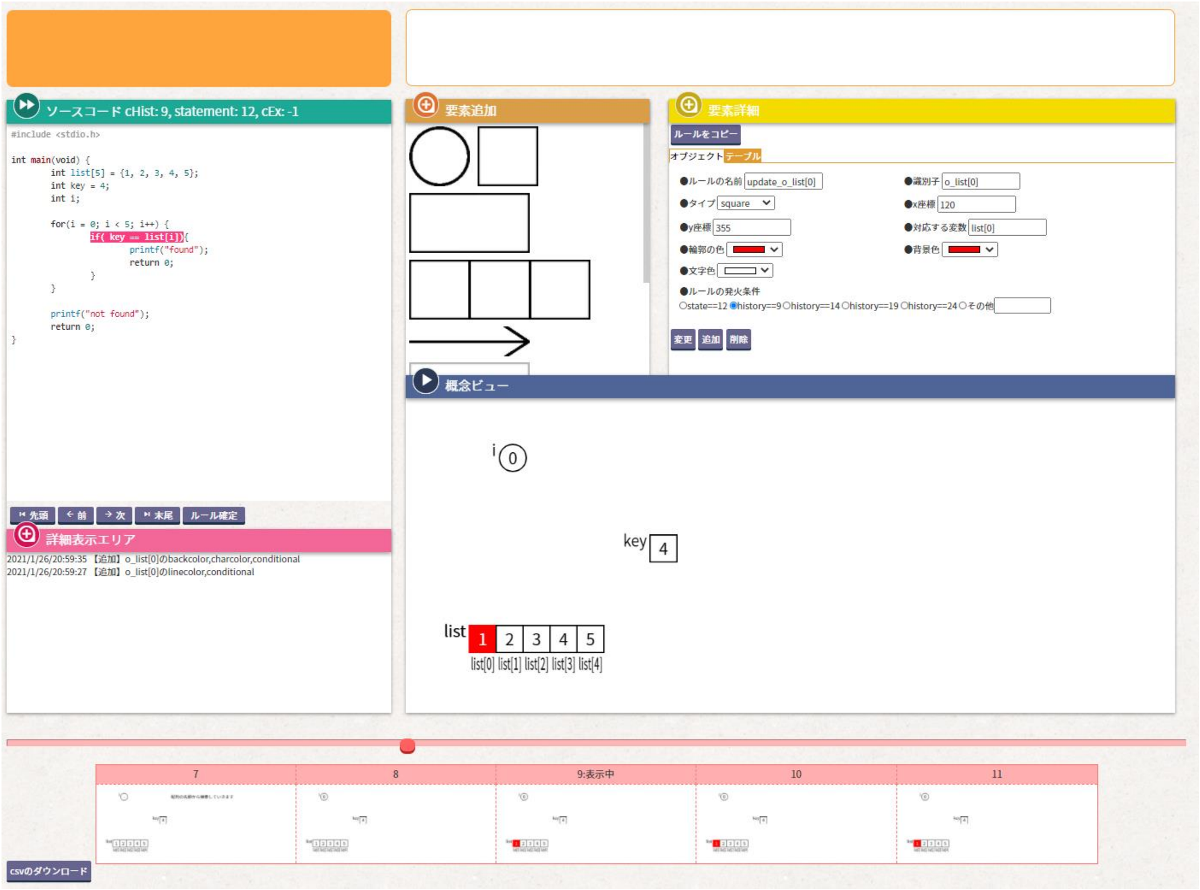
However, Tezuka GUI and existing WYSIWYG PV editors only support PV drawing. In general, PV systems change the drawing content along with the program-execution process. This allows learners to understand the function of each statement in the target program by observing the differences between PVs. Hence, we regard PVs as visualizations of the target domain world. The meaning of each statement in the program is defined by the extent to which executing the statement changes the target domain world. Importantly, PVs are not simply drawings of data structures but sequences of drawings linked to program-execution processes.

PV dynamics are a direct representation of computer-program dynamics, which reveal the trajectory of changes in a computer’s internal state, such as data structures changed continuously. By showing these changes directly, dynamic visualizations can offload a learner’s cognitive working memory, potentially enabling deeper cognitive processes. Dynamic visualizations can also facilitate cognitive processes that would otherwise require a lot of effort (Kühl, Scheiter, Gerjets, & Edelmann, 2011; Schnotz & Rasch, 2005). In other words, the learning effect of PV systems can be attributed to their dynamism. However, PV editors only support the creation of static visualizations. This study argues that PV creation cannot be fully supported through PV drawings alone. Instead, a function that helps capture a time-series sequence of PVs is required.

Based on this consideration, we aim to support PV creation more effectively by developing a GUI system that includes PV time-series information.

3. GUI System for PV Creation with Time Series Information

Figure 4 presents a screenshot of our developed GUI system, implemented in JavaScript. The *integrability* described in Section 2.1 includes an easy-installation feature.



ソースコード cHist: 9, statement: 12, cEx: -1

```
#include <stdio.h>

int main(void) {
        int list[5] = {1, 2, 3, 4, 5};
        int key = 4;
        int i;

        for(i = 0; i < 5; i++) {
                if( key == list[i]){
                        printf("found");
                        return 0;
                }
        }

        printf("not found");
        return 0;
}
```

要素詳細

オブジェクト テーブル

●ルールの名前 update_o_list[0]
●識別子 o_list[0]
●タイプ square
●x座標 120
●y座標 355
●対応する変数 list[0]
●輪郭の色 🟥
●背景色 🟥
●文字色 □
●ルールの発火条件
○state==12 ●history==9 ○history==14 ○history==19 ○history==24 ○その他

詳細表示エリア

2021/1/26/20:59:35 【追加】o_list[0]のbackcolor,charcolor,conditional
2021/1/26/20:59:27 【追加】o_list[0]のlinecolor,conditional

Figure 4. Screenshot of our GUI System.

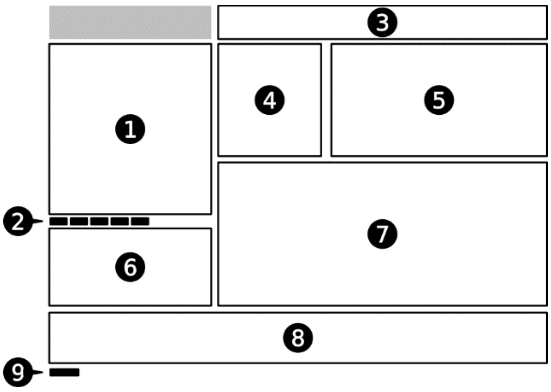


Figure 5. Nine Display Areas in our GUI System.

Our GUI system consists of nine display areas, as shown in Figure 5. The contents of each display area are as follows.

1. The program-code area displays the source code of the target C program to be visualized.
2. The control-button area enables PV creators to have stepwise control of target-program execution.
3. The message area displays operation messages to guide PV creators, for example, by notifying them of insufficient drawing-rule definitions.
4. The drawing-object area is used by PV creators to select a drawn object to be visualized on the PV.
5. The attribute area displays a list of attributes of the selected drawn object, allowing PV creators to set the attribute values using pull-down menus, combo boxes, and input forms.
6. The editing log area displays the editing history of the drawn object and its attribute values.

7. The PV area visualizes the PV by interpreting current drawing-rule definitions.
8. The seek bar and PV thumbnail area displays the current PV and those before and after it in thumbnail style. The PV creator can change the current PV by dragging the seek bar.
9. The save button area enables PV creators to store the PV sequence to a configuration file as a set of drawing rules.

The seek bar and PV thumbnails in Area 8 mainly represent the time-series information described in the previous section. PV creators can arbitrarily change the current PV within the time-series PV sequence by dragging the seek bar. PV thumbnails can make creators aware of the continuity of PVs. They also visually confirm the differences between adjacent PVs in the time series. The seek bar not only helps PV creators navigate the execution process but also helps them grasp the approximate position of the current PV in the time series. We intend to improve PV-generation efficiency using this feature.

4. Evaluation

To evaluate the effectiveness of PV creation using our GUI system, we conducted an experiment to measure the actual time needed to create or modify PVs. A survey by Naps et al. (2002) found that more than 90 percent of participants at the ITiCSE 2002 conference of the ACM cited the time required for PV creation as a factor in their reluctance to use animation (i.e., PV sequences). Thus, the evaluation of effortlessness, based on the time required for PV creation, is considered to be valid. The present study has measured the time required for PV creation using Tezuka GUI and evaluated the degree of improvement in our system's effortlessness.

Ten participants were involved in this experiment: two teachers with experience of teaching programming, five students with experience as programming teaching assistants, and three students with the same level of programming experience as the teaching assistants. We prepared two sample programs as the PV creation targets: a linear-search program and a maximum-value derivation program. Each participant received one sample program and was asked either to create a PV or to modify the PV provided. To reduce the order effects, we specified whether the participants would use Tezuka GUI or our system first. Table 1 briefly summarizes the conditions for each participant.

Table 1. *Conditions for Each Participant*

Participant #	Target	Operation	First use	Second use
1	Linear search	Creation	Tezuka GUI	Our system
2	Linear search	Modification	Tezuka GUI	Our system
3	Maximum value	Creation	Tezuka GUI	Our system
4	Maximum value	Modification	Tezuka GUI	Our system
5	Linear search	Creation	Our system	Tezuka GUI
6	Linear search	Modification	Our system	Tezuka GUI
7	Maximum value	Creation	Our system	Tezuka GUI
8	Maximum value	Modification	Our system	Tezuka GUI
9	Linear search	Creation	Tezuka GUI	Our system
10	Maximum value	Creation	Our system	Tezuka GUI

We began this experiment by explaining to the participants (for 45 minutes) the specification of TEDViT drawing rules and how to use the two systems. Next, we gave them a sample program and a sample PV, without disclosing the drawing rules, and asked them to define drawing rules to reproduce the sample PV. Participants assigned to modification received a set of drawing rules for the sample PV, which included some errors. We measured the time each participant took to define the appropriate drawing rules. Subsequently, we conducted a questionnaire survey on the effectiveness of the seek bar and PV thumbnail function, using a five-point grading system. We also conducted brief interviews to ascertain participants' opinion regarding the two systems.

Table 2 presents the experimental results. Regardless of the order in which they used the GUI systems, target programs, and task operations, all participants took less time to complete the task with our system than with Tezuka GUI. The reduction rate, based on the average time spent by all participants, was 41.3%. This suggests that the PV creation with our GUI system significantly improves effortlessness. However, this experiment compared two full systems, without specifically evaluating the effectiveness of time-series information (the main focus of this study). To obtain more precise results, we must develop two systems, one with a seek bar and PV thumbnail functions and the other without, and measure the time required for PV creation in both systems. We plan to conduct such experiments in the future.

Table 2. *Measured Times for PV Creation/Modification*

Participant #	Time with Tezuka GUI (sec)	Time with our system (sec)
1	1136	925
2	421	249
3	1935	989
4	1360	320
5	914	834
6	523	416
7	1355	845
8	520	377
9	2022	1015
10	1383	819

The questionnaire survey on the effectiveness of the seek bar and PV thumbnail function produced an average score of 4.2, suggesting that the participants gave the function a positive rating. In the interview survey, some participants commented that providing PV continuity alongside the time series made it easier to spot errors. Dividing the task times into two groups, based on participant operations, reduced the average times needed for PV creation and modification by 37.9% and 51.8%, respectively. This suggests that our GUI system supports the task of correcting errors more effectively than Tezuka GUI. This finding will be useful in future evaluations of ways to improve effortlessness in PV creation.

5. Conclusion

The system described in this paper attempts to improve effortlessness in PV creation by incorporating time-series information into a GUI PV-creation system.

One of the main obstacles to the continuous introduction of PV systems in actual classrooms is the significant amount of time needed to integrate PV systems into actual educational setups. Our previous study has developed a PV system called TEDViT and introduced it into several practical classes. While programming learning with TEDViT has a certain effect, the time required for PV customization is a non-trivial problem. Using a GUI-based WYSIWYG PV editor is a promising way to address this issue. However, many existing systems support only the drawing of PVs, even though PVs are not simply drawings of data structures but sequences of drawings alongside a program-execution process. This study has therefore developed a PV-creation support system that takes into consideration the continuity of drawings by incorporating time-series information into the GUI. We conducted an evaluation experiment to measure the time needed to create PVs using Tezuka GUI and our system. The results showed that our GUI system reduced the average time by 41.3%, when compared to Tezuka GUI. This suggests that our GUI system improves the efficiency of PV creation.

The experiment in this study evaluated the effortlessness of the entire GUI system. In future research, we will conduct experiments that focus specifically on the effectiveness of time-series information. We plan to measure the difference between PV-creation times using systems with and without the seek bar and thumbnail function to reflect time-series information. As the participants in the present experiment subjectively felt that the seek bar and thumbnail function had a positive effect on PV creation, we expect to obtain positive results. The results also reveal that our system is more effective than Tezuka GUI in correcting PVs that contain errors. This will be a useful finding as we consider future effortlessness improvements.

Acknowledgements

This study was supported by JSPS KAKENHI Grant Numbers JP18K11566, JP18K11567, and JP19K12259.

References

Guo, P. J. (2013). Online Python tutor: Embeddable web-based program visualization for CS education. *Proceedings of the 44th ACM Technical Symposium on Computer Science Education (SIGCSE)*, 579–584.

Helminen, J., & Malmi, L. (2010). Jype—A program visualization and programming exercise tool for Python. *Proceedings of the 5th International Symposium on Software Visualization*, 153–162.

Ihantola, P., Karavirta, V., Korhonen, A., & Nikander, J. (2005). Taxonomy of effortless creation of algorithm visualizations. *Proceedings of the First International Workshop on Computing Education Research*, 123–133.

Karavirta, V., Korhonen, A., Nikander, J., & Tenhunen, P. (2002). Effortless creation of algorithm visualization. *Proceedings of the Second Annual Finnish/Baltic Sea Conference on Computer Science Education*, 52–56.

Kogure, S., Fujioka, R., Noguchi, Y., Yamashita, K., Konishi, T., & Itoh, Y. (2014). Code reading environment according to visualizing both variable's memory image and target world's status. *Proceeding of the 22nd International Conference on Computers in Education (ICCE2014)*, 343–348.

Kühl, T., Scheiter, K., Gerjets, P., & Edelmann, J. (2011). The influence of text modality on learning with static and dynamic visualizations. *Computers in Human Behavior*, *27*(1), 29–35.

Malone, B., Atkinson, T., Kosa, M., & Hadlock, F. (2009). Pedagogically effective effortless algorithm visualization with a PCIL. *Proceedings of the 39th IEEE International Conference on Frontiers in Education Conference*, 1501–1506.

Moreno, A., Myller, N., Sutinen, E., & Ben-Ari, M. (2004). Visualizing programs with Jeliot3. *Proceedings of the Working Conference on Advanced Visual Interfaces*, 373–376.

Naps, T. L., Rößling, G., Almstrum, V., Dann, W., Fleischer, R., Hundhausen, C., Korhonen, A., Malmi, L., McNally, M., Rodger, S., & Velázquez-Iturbide, J. (2002). Exploring the role of visualization and engagement in computer science education. *Working Group Reports from the 2002 Conference on Innovation and Technology in Computer Science Education*, 131–152.

Pears, A., Seidman, S., Malmi, L., Mannila, L., Adams, E., Bennedsen, J., Devin, M., & Paterson, J. (2007). A survey of literature on the teaching of introductory programming. *ACM SIGCSE Bulletin*, *39*(4), 204–223.

Rößling, G., & Ackermann, T. (2007). A framework for generating AV content on-the-fly. *Electronic Notes in Theoretical Computer Science*, *178*, 23–31.

Rößling, G., & Freisleben, B. (2002). ANIMAL: A system for supporting multiple roles in algorithm Animation. *Journal of Visual Languages & Computing*, *13*(3), 341–354.

Schnotz, W., & Rasch, T. (2005). Enabling, Facilitating, and Inhibiting Effects of Animations in Multimedia learning: Why reduction of cognitive load can have negative results on learning. *Educational Technology Research and Development*, *53*(3), 47–58.

Sirkiä, T., & Sorva, J. (2015). Tailoring animations of example programs. *Proceedings of the 15th Koli Calling Conference on Computing Education Research*, 147–151.

Sorva, J., Karavirta, V., & Malmi, L. (2013). A review of generic program visualization systems for introductory programming education. *ACM Transactions on Computing Education (TOCE)*, *13*(4), 1–64.

Tezuka, D., Kogure, S., Noguchi, Y., Yamashita, K., Konishi, T., & Itoh, Y. (2016). GUI based environment to support writing and debugging rules for a program visualization tool. *Proceedings of the 24th International Conference on Computers in Education (ICCE2016)*, 303–305.

Tudoreanu, M. E. (2003). Designing effective program visualization tools for reducing user's cognitive effort. *Proceedings of the 2003 ACM Symposium on Software Visualization*, 105–114.

Velázquez-Iturbide, J. Á., Pareja-Flores, C., & Urquiza-Fuentes, J. (2008). An approach to effortless construction of program animations. *Computers & Education*, *50*(1), 179–192.

Yamashita, K., Fujioka, R., Kogure, S., Noguchi, Y., Konishi, T., & Itoh, Y. (2016). Practice of algorithm education based on discovery learning using a program visualization system. *Research and Practice in Technology Enhanced Learning (RPTEL)*, *11*(15), 1–19. doi:10.1186/s41039-016-0041-5

Yamashita, K., Fujioka, R., Kogure, S., Noguchi, Y., Konishi, T., & Itoh, Y. (2017). Classroom practice for understanding pointers using learning support system for visualizing memory image and target domain world. *Research and Practice in Technology Enhanced Learning (RPTEL), 12*(17), 1–16. doi:10.1186/s41039-017-0058-4

Yan, Y., Nakano, H., Hara, K., Suga, S., & HE, A. (2014). A C programming learning support system and its subjective assessment. *Proceedings of 2014 IEEE International Conference on Computer and Information Technology*, 561–566.

Co-Designing for a Healthy Edtech Ecosystem: Lessons from the Tulna Research-Practice Partnership in India

Aastha PATEL[a*], Chandan DASGUPTA[a], Sahana MURTHY[a] & Rashi DHANANI[b]
[a]*Indian Institute of Technology, Bombay, India*
[b]*Central Square Foundation, India*
*aasthapatel@iitb.ac.in

Abstract: The demand and supply of EdTech products have surged in the past decade and especially post-Covid19. Yet key challenges exist, such as inadequate quality standards and lack of reliable product evaluations. Consequently, stakeholders in the EdTech ecosystem such as schools, teachers, parents, governments, philanthropists, and investors face difficulty in making informed adoption decisions and feel the need for a systematic quality evaluation framework of educational technology (EdTech) products. In such a scenario, we analyze a Research-Practice Partnership between educational researchers, government decision-makers, and a non-governmental organization working on policy and strategy, who collaborated on designing and implementing EdTech quality standards. We examine the co-design process of the 'Tulna EdTech evaluation index' at various stages in the partnership. We adopted the Design-based Implementation Research (DBIR) approach for guiding our partnership. We examine what can be learned from the process of co-design in our RPP that might be useful for our ongoing partnership going forward, as well as for other RPPs. We found that the stakeholders navigated through tensions and iteratively negotiated the design of the evaluation index based on their individual expectations, perspectives, and expertise. Our retrospective, qualitative analysis supports understanding of how researchers and practitioners might engage in co-design and the role co-design might play in establishing a healthy EdTech ecosystem.

Keywords: EdTech in India, Design-based Implementation Research (DBIR), EdTech quality standards, Co-Design, Research-Practice Partnership

1. Introduction

Today's educational settings all over the world have seen the influx of several hundreds of educational technology (EdTech) products. The EdTech market is growing rapidly, even in developing countries. The demand for EdTech is also high and is driven by governments, private organizations, schools as well as individual decision-makers such as parents. The variety of EdTech based solutions is enormous in terms of the intended goals, target audience, technologies, pedagogical design, cost, use case, and so on. At the same time stakeholders in the EdTech ecosystem face key challenges in understanding what constitutes good quality EdTech. Research studies have given rise to recommendations for the design and implementation of EdTech solutions, for instance on addressing goals such as personalization through appropriate pedagogical strategies and technology affordances (Peng, 2019), or on scaffolding students for various learning needs (Quintana et al., 2004). There exist a few evaluation frameworks and instruments (LORI; Nesbit & Leacock, 2007; LOESS; Kay & Knaack, 2009) but are generic across different types of EdTech products, may not have been tested for validity and reliability, and often have not been implemented with products on the market.

In developing countries, this problem of practice is further exacerbated due to inadequate quality standards that take into account the needs arising from the local EdTech ecosystem, and a lack of unbiased reliable product evaluations. In economic terms, this information asymmetry leads to ad hoc and inefficient decision making for adoption, without regard to the impact on learning (Ferreyra, M., & Liang, P., 2012) In such scenario, stakeholders such as schools, teachers, parents, governments, philanthropists and investors feel the need for a systematic quality evaluation framework of EdTech products to make informed decisions.

The problem of lack of EdTech quality standards is systemic in nature involving multiple stakeholders in their individual roles and collective roles. To address this issue, a solution needs to be designed that actively engages all the stakeholders and entities involved who need to not only align but also collaborate. Such a collaborative, systemic effort to design and implement quality standards contribute towards creating a healthy EdTech ecosystem (Omidyar Network, 2019). For a healthy ecosystem to emerge and sustain, the solution needs to integrate the requirements and practices of the various stakeholders. For example, the solution should include research-based quality standards that are robust as well as usable. It should encourage demand for evidence-based recommendations and also garner acceptance from the product supply perspective.

One model of such integration is a co-design process within a Research-Practice Partnership (RPP). RPPs enable the use of research evidence in decision-making for policymakers and funders (Tseng, 2012). In this paper, we describe an RPP formed as part of the *EdTech Tulna* initiative in India (Tulna, 2021). The partnership is between an EdTech research group from a research university, a non-governmental organization working on policy and strategy, and a government agency involved in EdTech procurement decisions. A design-based implementation research (DBIR) approach (Fishman et al., 2013) guided us in the co-design of the 'Tulna EdTech evaluation index' at various stages in the partnership. We unpack the co-design process from the lens of RPP and DBIR and present a case study involving an analysis of three co-design episodes. The lessons learned from this RPP illustrate how researchers and practitioners might engage in co-design and work towards establishing a healthy EdTech ecosystem in India.

2. Background and Context

2.1 EdTech Ecosystem in India

The past two decades have seen a strong emergence of technology-enhanced learning in India in all sectors. Large-scale initiatives by the Government of India such as the National Mission on Education through ICT (NMEICT, 2021) focused on access, quality and equity through providing connectivity and devices as well as on content generation. In addition, traditional academic publishing houses reoriented towards a technology division, adding repositories of ICT-based content. Non-governmental organizations (such as Central Square Foundation (CSF, 2021)) and international foundations have devoted tremendous efforts towards improving the school education system, especially in low-income communities with an emphasis on technology-supported solutions for scaling and access. Investors (for example, Omidyar Network (2019)) look to EdTech based solutions to address the problem of quality education. In addition, in recent years, the market has witnessed the influx of several hundred EdTech products from for-profit product companies as well as not-for-profit organizations funded by philanthropists and foundations. The National Educational Policy (NEP, 2020) emphasizes technology use and integration by establishing a 'National Educational Technology Forum' that will advise the leadership of educational institutions and facilitate decision-making based on induction, deployment, and use of EdTech.

Within such a dynamic scenario, adoption decisions are a key challenge. Government decision-makers are not equipped to differentiate between various types of EdTech and to examine their quality from different perspectives. It is also time-consuming for government panels to conduct proof of concept evaluations. Thus what typically gets emphasized is infrastructure, cost, or convenience, and what gets left out are parameters related to learning and teaching. Reputed product companies value unbiased and credible evaluations both as a pathway to adoption as well as a formative feedback mechanism to improve the product. A prior study of 12 EdTech companies in India has shown the need for public evaluations and a systematic approach to evaluate a large number of products on a regular basis (CSF EdTech Lab 1 Report, 2019). Individual decision-makers such as parents are faced with an array of products. They have to navigate these choices and make financial decisions mostly on the basis of word-of-mouth recommendations, while at the same time facing strong advertising from vendors.

In mature EdTech markets, various models of evaluation and quality frameworks exist (for example EdReports (EdReports, 2021)). Such frameworks, standards, or evaluations that are relevant to the local context are not yet part of the Indian EdTech ecosystem.

2.2 EdTech Tulna Initiative and the Partners in The RPP

The *EdTech Tulna* initiative (Tulna, 2021) in India has been created as a public good to address the challenges of information asymmetry around EdTech quality. *EdTech Tulna* consists of a research-based framework to set quality standards and a corresponding product evaluation index. The index is applied to evaluate existing products in the ecosystem and the evaluations are intended for public use. The aim is to promote demand for evidence-based decision-making and encourage the supply of interventions to meet the demanded quality standards. EdTech Tulna index focuses on evaluating the design of EdTech products along three constructs: Content Quality, Pedagogical Alignment, and Technology & Design to capture a holistic view of the quality of the product design. Each construct further comprises multiple criteria.

EdTech Tulna is created out of a partnership in India between the Educational Technology department at IIT Bombay, a premier research institute in India, and the Central Square Foundation, a non-governmental organization focusing on educational policy and strategy. The RPP discussed in this paper has another partner, a state government that is involved in making decisions for EdTech adoption at scale. The state government, along with a team of consultants that facilitate product procurement and adoption are considered as the practitioners in the RPP. This triadic partnership model brings a wide range of knowledge, skills, and practices that are essential for addressing the EdTech quality problem. Fig. 1 depicts the areas of expertise and the roles of individual partners in this RPP as well as those emerging at the collaboration interface.

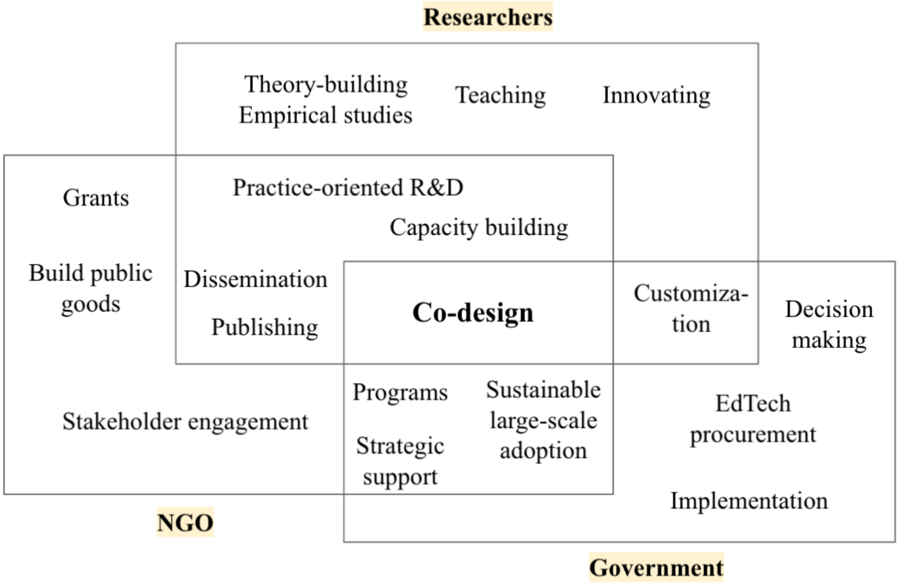


Figure 1. Co-design emerging from the Triadic Partnership between Researchers, NGO and Government.

The researchers have an academic background and engage in research involving theory-building and empirical studies. A key area of practice for academic researchers is teaching and training students in disciplinary knowledge and skills. Through these activities, academic researchers are engaged in discovering new knowledge and innovating solutions. For the NGO, the impact areas of focus are foundational learning, bringing innovation in education through technology-based solutions, and improving learning outcomes at scale for low income children. Towards these, they engage with other stakeholders in the ecosystem to build public goods, run programs and create research-based solutions. They also support organizations through grants to drive sustainable impact in these areas. The government's role in this context includes making decisions regarding EdTech adoption for large numbers of schools within its jurisdiction and implementing the processes for procurement. The collaboration between the researchers and the NGO has given rise to practice-oriented research and development projects with an emphasis on data collected from field implementations. It has also led to capacity-building efforts and the publication of research reports and guidelines for various stakeholders. For this project, the NGO helped the researchers better understand the context for the design. The NGO collaborates with the government and the consultants in providing strategic and technical support along with the management and implementation of programs, with a focus on scaling and sustainability. The collaboration interface between the researchers and the government consultants in this RPP was a novel exercise in the customization of a research tool to address the requirements of the government's policies and procedures. What emerged at the heart of this three-way partnership was an effort in co-designing a solution for EdTech quality evaluation that drew from the diverse expertise and addressed diverse needs. This partnership includes educators who have experience working in educational institutions but we have not formally engaged with schools as partners yet due to logistical challenges.

3. Theoretical framework

Research practice partnerships (RPPs) are long-term collaborations between practitioners and researchers that are organized to investigate problems of practice and solutions for improving system outcomes. There are multiple types of RPPs such as Research Alliances, Network Improvement Communities, and Design Research Partnerships (Coburn, Penuel, & Geil, 2013). Our partnership is categorized as a 'Design research partnership' as it focuses on informing practice and research by designing, developing, and testing an educational innovation - EdTech Tulna - an EdTech Evaluation Index. Design Research partnerships typically have the following characteristics. These are place-based, have dual goals of informing practice and research, emphasize co-design, and rely on close collaboration at every stage in the process (Coburn et al., 2013). The principle of co-design is central to this type of partnership. By co-design we refer to "a highly facilitated process that engages people with diverse expertise (e.g., research, curriculum, professional development, teaching) in designing, developing, and testing innovations." (Coburn et al., 2013, p.10).

Design research partnerships are often formed using the Design-Based Implementation Research approach (DBIR) (Fishman et al., 2013) since this approach caters to both design and implementation of innovative interventions involving multiple stakeholders. We use DBIR as a framework to analyze the co-design process. The core principles that characterize DBIR are:

- *Principle 1*: Teams form around a focus on persistent problems of practice from multiple stakeholders' perspectives.
- *Principle 2*: To improve practice, teams commit to iterative, collaborative design.
- *Principle 3*: To promote quality in the research and development process, theory guides both learning and implementation through systematic inquiry.
- *Principle 4*: DBIR is concerned with developing capacity for sustaining change in systems.

These principles guide our partnership and help make our co-design process more reliable in terms of the outcomes (Penuel, 2019).

4. Method

We draw anecdotes from the ongoing Tulna RPP focused on establishing EdTech Quality standards in India. The data were collected over a period of 8 months and documented systematically in the form of meeting notes, documents, email communication, versions of designed artifacts, and participant observation. Meetings were conducted twice a week at various levels such as internal research meetings, partner meetings, and meetings with external stakeholders. Due to covid, all meetings happened virtually and collaboration was completely remote. We analyzed notes and transcripts from the meetings. In the first round of analysis, we identified the *episodes of co-design* where multiple stakeholders were involved in creating something emergent. Three main episodes of co-design emerged from our analysis and they were selected based on their chronological occurrence and stakeholders involved in co-design. One episode each was selected from the three phases: (i) Project Scoping (ii) Development of the Evaluation index and (iii) Customisation of the index. In the second round of our analysis, we analyzed each episode using the core principles of DBIR. The analysis was conducted by the research team, the authors of the study were the active members in the co-design process, and presented the first-person account of co-design (Coburn & Penuel, 2016). In the following section, we first describe the three episodes, highlighting what each stakeholder brought for co-design and how it led to key emergent developments in the partnership.

5. Case Study: Co-Design between Educational Researchers, Non-Governmental Organization, and Government Consultants

5.1 Co-Design Episode 1: Systematic Layered Approach to EdTech Evaluation

Aligned to the first guiding principle of DBIR, the Tulna partnership between researchers and the policy NGO brought in multiple stakeholders' perspectives to solve the pressing problem of the lack of quality standards in EdTech. The partners aligned on the broad problem of practice and further carefully identified root causes, key change drivers to develop a long-term vision, and practical theory of action. At this stage, partnership often gets challenging as researchers and practitioners work at different scales. Researchers start by investigating small discrete elements in controlled settings with few users whereas practitioners are likely to use those measures to investigate the problem at scale in-situ with many users (Tseng, 2012). In this case, the research team preferred to focus on the fundamental design of EdTech products and defining quality standards at a product level, since research indicates that a product can succeed at scale only if it is designed well. On the other hand, the NGO's emphasis was on evaluating the product at scale, taking into consideration infrastructural issues that may affect the use of the product on the ground. Both these perspectives are valid in the case of evaluating EdTech products. However, negotiating these multiple perspectives held by the two stakeholders was important in this partnership in order to make progress. Both the research team and NGO worked together to determine the scope of the project that might allow both perspectives to co-exist. Numerous meetings were held between these two stakeholders to jointly design what this evaluation approach might look like. Both the stakeholders were also trying to negotiate the timeline by which the evaluations would be done. Research suggests that there is a difference in researcher's and practitioner's timelines of work, Practitioners often work under stringent timelines for immediate implementations of innovations, whereas researchers proceed slowly with relatively flexible timelines through cycles of inquiry and analysis before they are ready for recommending action (Penuel & Allen, 2015). The dual negotiation points of what to evaluate and when to evaluate informed the joint decisions taken by the stakeholders.

A layered approach to quality evaluation (Fig. 2) emerged from these joint discussions. This approach enabled multiple perspectives to co-exist and presented a new roadmap for the partnership that had not existed before. As per this approach, the research team started the evaluation process by focusing on the design of products, i.e. the foundational layer in Fig. 2. The long-term evaluation plan now also consisted of two intermediate layers - evaluating the interaction of learners and teachers with the product (curated user studies) and evaluating learning outcomes in the field setting (in-situ studies). The final stage of the evaluation program comprised evaluation of the outcomes at scale.

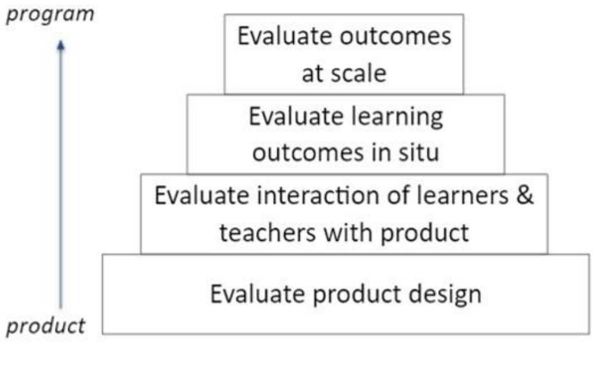

program
Evaluate outcomes at scale
Evaluate learning outcomes in situ
Evaluate interaction of learners & teachers with product
Evaluate product design
product

Figure 2. A Systematic Layered Approach to Evaluation.

This approach was accepted by both partners and helped develop a long-term vision for sustainable outcomes. The rationale provided by the researchers of understanding both the outcomes and the underlying mechanisms convinced the practitioners to come on board, and helped build confidence that all aspects will be addressed but not just immediately. This also helped the partners navigate the different scales and time frames between research and practice (Tseng, 2012).

5.2 Co-Design Episode 2: Re-Defining the 'Contextualization' Construct

This episode is aligned with the third principle of DBIR and describes the co-design process for developing the evaluation index. In the first version of the index, the researchers had included 'contextualization' as one of the constructs for evaluating the EdTech products. Contextualization was defined as a measure of support provided to the teachers and learners to enhance its potential effectiveness in the users' local context. The need for the idea of contextualization in EdTech products is agreed upon by all stakeholders in the ecosystem. This becomes especially important for India because of its diversity in context across different states, 100+ languages spoken in the country, and the disparity in socio-economic classes. However, discussions with the NGO soon made it clear that different stakeholders interpreted contextualization differently. The NGO conveyed that 'contextualization' is a commonly used terminology in large-scale implementation programs where practitioners conduct pilot studies for contextualization before scaling programs. The research team argued for including it as an evaluation construct from the pedagogical, content, and interaction point of view where each product's design needed to be contextually relevant for the target teachers and students. While the NGO agreed with this point of view, it considered infrastructural issues and training and support available for teachers and students as more important factors determining the effectiveness of a product in a context. The research team acknowledged that these factors were relevant for evaluating a product for contextualization. Along with the multiple interpretations of contextualization that emerged from the discussions with the NGO, the researchers also realized that it is not possible to evaluate contextualization entirely from the product's design. One needs to conduct user studies (top three layers in Fig 2) to evaluate a product for contextualization - from the pedagogical, content, and interaction perspective as well as infrastructural and training perspective.

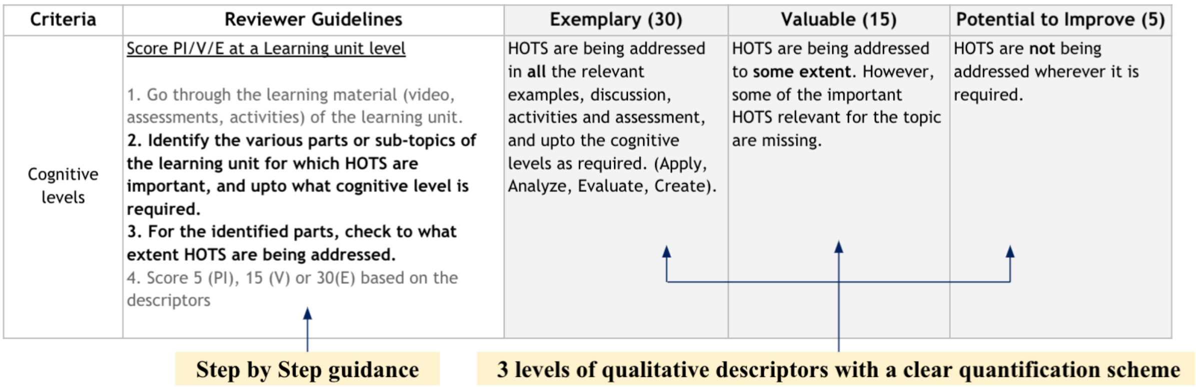
After further negotiations, it was agreed that there were some aspects of contextualization that could be evaluated within the product design. The researchers reviewed the literature to identify these which were then absorbed under the other three constructs of the Tulna index - Content quality, Pedagogical alignment, and Technology & design. Thus, these three constructs were co-designed to now cater to a broader scope that covered contextualization. Aligning to the third principle of DBIR, both partners argued for the extent to which contextualization should be included in the index and the systematic layered approach to evaluation (Figure 2) gave them the flexibility to push it to the upper layers of evaluation i.e., later on in the program. This high-level co-design between the partners evolved the index to be more robust and univocal.

5.3 Co-Design Episode 3: Quantification of the Index

This co-design episode aligns with the second and fourth DBIR principles. Once the Tulna index was in the advanced stages of development, the NGO facilitated interactions with the state government through a group of government consultants. This was the first instance of the adoption of the index by a state government. The Tulna index was considered as a comprehensive tool to evaluate the quality of EdTech products for the state's large-scale EdTech adoption for 2000+ schools. In the initial phases, the government consultants conveyed the criteria that they considered important for the evaluation of EdTech products in their setting. The research team took these criteria into consideration while designing and validating the index. However, when this adoption process began, all the three partners realized that the government needed an index that i) enabled rapid decision-making, and ii) was capable of providing numerical scores. Several rounds of discussions between the partners highlighted a tension between research rigor captured in the index and the need for ease of use by government officials. The original version of the index was designed to be used by trained evaluators. All partners realized that the index was not suitable as is to be used by government officials and in order to make it usable they had to customize it.

Aligning with the second guiding principle of DBIR the researchers and the government consultants collaboratively adapted the index to a more usable and contextualized format for this specific state government. This process of partnership went beyond the usual division of labor where the researcher would design the index, while practitioners would test and implement it (Coburn & Penuel, 2016). The government consultants, as well as the NGO, played a vital role in bringing out the need for quantification and suggested the researchers assign numerical values to the existing 3-scale descriptive rubric along with a comprehensible numerical target descriptor. For example, they requested to rephrase *"Higher Order Thinking Skills (HOTS) are being sufficiently addressed in the content, examples, activities"* to *"80% or more HOTS are being addressed in the content, examples, activities".* The research team conveyed to the other stakeholders that quantification at this level was not feasible as there was no basis for claiming that a certain percentage of content was sufficient to make a product effective. Further, the government and the NGO preferred an index that represented the evaluation of an EdTech product in the form of one numerical value for easy decision-making, rather than a descriptive qualitative rubric. The researchers understood the need for quantification but emphasized that a qualitative rubric enables bringing out pros and cons in the product's design that would be lost in a quantified number.

Negotiations between the government consultants and the research team resulted in the research team agreeing to assign numerical values in a way that helps the consultants synthesize and make decisions. A quantitative scoring scheme was devised having three nonlinear scales i.e. Exemplary (score 30), Valuable (score 15), and Potential to Improve (score 5). This nonlinear scaling came from a research perspective. Thus, in order to bring structure and objectivity to this complex process, a quantified version of Tulna index was created for this state government. The quantified version of the Evaluation index was designed keeping in mind the practitioners as the primary users of the index. Further, clearly articulated reviewer guidelines for step-by-step guidance for each evaluation criterion were added in this quantified version of the index. Supplementary material was provided to aid the understanding of disciplinary concepts and also help navigate through the index. In order to reduce the cognitive overload of the government officials, important steps were emphasized using black bold text and some tacit guidelines were explicitly mentioned in grey text for anyone who needed guidance at a granular level. Fig. 3 illustrates the characteristics of the quantified version for one criterion.



Criteria	Reviewer Guidelines	Exemplary (30)	Valuable (15)	Potential to Improve (5)
Cognitive levels	Score PI/V/E at a Learning unit level 1. Go through the learning material (video, assessments, activities) of the learning unit. **2. Identify the various parts or sub-topics of the learning unit for which HOTS are important, and upto what cognitive level is required.** **3. For the identified parts, check to what extent HOTS are being addressed.** 4. Score 5 (PI), 15 (V) or 30(E) based on the descriptors	HOTS are being addressed in **all** the relevant examples, discussion, activities and assessment, and upto the cognitive levels as required. (Apply, Analyze, Evaluate, Create).	HOTS are being addressed to **some extent**. However, some of the important HOTS relevant for the topic are missing.	HOTS are **not** being addressed wherever it is required.

Step by Step guidance
3 levels of qualitative descriptors with a clear quantification scheme

Figure 3. Screenshot from the Quantified Evaluation Index.

At the end of the co-design process, the researchers realized the importance of quantification and ease of use in order to enable the adoption of the Tulna index by multiple state governments. The state government realized the importance of the multi-dimensionality of EdTech evaluations that could not be captured by a single numerical value for the entire product. The outcome of this co-design resulted in the design of a more usable, accessible, and practitioner-centered evaluation index. This new quantified version of the index can be easily carried forward into future practice and decision-making in other contexts. This aligns with the fourth guiding principle of DBIR. The researchers also conducted a short 2-day training workshop with the actual users of the index. They built on this experience to refine and develop a set of training resources that could be used in future collaboration with other state governments.

6. Discussion

In this paper, we have described and analyzed an ongoing design research partnership between a research team, NGO, and government consultants in India working towards defining EdTech quality standards, setting up the process and index for evaluating EdTech products, and enabling research-based decision-making by the stakeholders. The long-term vision shared by all stakeholders is to establish a healthy EdTech ecosystem in India that is characterized by demand for and supply of good quality EdTech products. We used the DBIR approach to describe what this partnership looked like using the three representative episodes. Co-design forms the core of this partnership (Coburn et al. 2013). As highlighted in the literature, this co-design process was place-based i.e. influenced by the state government (represented by the government consultants) that wanted to first adopt the Tulna index and informed the creation of the quantitative version of the index. The process informed research and practice. The research team developed a better understanding of the contextualization construct that was embedded within other evaluation constructs. The NGO and government understood the importance of these evaluation constructs containing contextualization. The state government's evaluation process for procuring EdTech products also changed as a result of this partnership. They used Tulna index's multi-dimensional evaluation criteria for evaluating products' design during the procurement process. Instead of using a single number for describing the quality of the product, they now considered quantitative scoring across the three dimensions of content quality, pedagogical alignment, and technology and design. This partnership has been marked by numerous episodes of close collaboration between the stakeholders that resulted in defining a long-term roadmap for the evaluation program consisting of a tiered approach starting with evaluating a product's design.

Specifically, our findings highlight the following lessons for sustaining co-design between multiple stakeholders aiming to establish a healthy EdTech ecosystem in India.

Lesson 1: Leverage multiple interpretations

Co-design involves numerous instances of multiple interpretations due to differences in the stakeholder's goals and cultural influences defining the way the organization works. In our partnership, we observed multiple perspectives about the desired approach for evaluating EdTech products, contextualization as an evaluation dimension, and the need for quantification of evaluation results. Such episodes became the anchor points for our co-design process. Each stakeholder brings with them a unique perspective and it is essential to enable these perspectives to co-exist from the very beginning of the partnership. In RPPs, it is essential to bring diverse perspectives together in productive ways while maintaining mutualism (Coburn et al., 2013). Both researchers and practitioners must acknowledge that differences are bound to exist but leveraging these differences will likely lead to a collaborative endeavor where the sum is bigger than the combination of the parts contributed by each stakeholder. For instance, in the first co-design episode, multiple interpretations of the approach to evaluation resulted in the creation of a layered approach that presented a new roadmap for the partnership that had not existed before. In the second episode, the stakeholders had different interpretations of contextualization as an evaluation construct. These gave rise to a refined evaluation index where key aspects of this construct were embedded within content quality, pedagogical alignment, and technology and design constructs for evaluating EdTech products. In the third episode, multiple perspectives for the need for quantification resulted in a multi-dimensional quantified version on the evaluation index that served the purpose of the state governments as well as researchers.

Lesson 2: Negotiate tension between research rigor and usability

Research-based educational interventions are guided by multiple disciplinary theories. Tulna index has deep roots in the theories of learning sciences, educational disciplinary knowledge, Human-Computer Interaction, and educational psychology. This may result in the creation of research-based tools that are usable only by researchers or highly trained individuals. This high entry barrier prevents the adoption of these tools by all the stakeholders. In RPPs, the dual goals of improving research rigor and pushing out research to practice (Tseng, 2017) lead to tension between research rigor and usability. Tension develops as researchers are trained to work towards high rigor and the practitioners are concerned about usability, effective adoption, and scaling of various tools. Prior research highlights one such tension between research rigor and timelines (Coburn et al. 2013). Rather than disrupting design, these design tensions can serve as opportunities for productive designs and they represent the need to be addressed in co-design (Severance et al., 2014; Tartan, 2007). Stakeholders need to negotiate this tension between the research rigor and practical usability. While all stakeholders may agree that both rigor and usability are important to have, the negotiation process should establish the extent to which each can be preserved. For instance in the third episode, while all stakeholders agreed that quantification was needed, quantification to the extent of specifying the percentage of presence of higher-order thinking skills (suggested by government stakeholder) was not possible because it was not supported by research. Instead, a quantitative scoring scheme was devised having three nonlinear scales i.e. Exemplary (score 30), Valuable (score 15), and Potential to Improve (score 5). In addition, detailed reviewer guidelines and evaluation descriptors that can help practitioners assign these numerical scores for decision-making were also designed. Capacity-building training programs were also conducted for the practitioners to help them use this quantified index. Similarly, in the first episode, the layered approach emerged as an outcome of negotiation between the need for a rigorous evaluation of a product's design and the need to evaluate the effectiveness of products at scale. In the second episode, the scope of the contextualization construct was reduced and merged with other constructs to make the evaluation index more usable by practitioners.

Lesson 3: Develop shared expectations

During the co-design process, stakeholders are likely to have a different understanding of each other's as well as their own roles in the partnership. We have outlined some of the unique and shared roles of the three stakeholders engaged in this co-design process in Fig. 1. The shared roles provide opportunities for developing shared expectations in the partnership. Prior research suggests that partnerships need to periodically revisit their shared understanding about their role and synthesize shared expectations and vision around the work that they are engaged in doing (Farrell, Harrison, & Coburn, 2019). The importance of this was reflected in the co-design process observed in our context as well. For instance, in episode 3, significant traction in terms of developing shared expectations was achieved between the researchers and the government agency while engaged in the shared work of customization and fine-tuning of the quantitative version of the Tulna evaluation instrument. Researchers drew on their role of theory-builders and domain experts and the government agency drew on their role involving EdTech product procurement to define the shared work around customization and quantification during the co-design process. The research team had begun with the expectation that only trained experts will be the evaluators using the Tulna index. However, with the new shared expectation of state agencies doing the actual evaluation of EdTech products quickly, the customization process gave rise to new tools and resources. Similarly, in Episode 2, shared expectations developed around the feasibility of data collection and evaluation using the contextualization construct. This process enabled the NGO to draw on its primary mandate and experience with stakeholder engagement and researchers to draw on their knowledge of theories of learning to give rise to the co-design process.

7. Conclusion

In this paper, we began with the objective of unpacking the co-design process and sharing lessons learned from an ongoing RPP to illustrate how researchers and practitioners might engage in co-design for establishing a healthy EdTech ecosystem in India. In the absence of examples of such partnerships in India, we had to create one from scratch. We have highlighted three lessons that emerged from this partnership, namely - leverage multiple interpretations, negotiate the tension between research rigor and usability, and develop shared expectations. These lessons are likely to help future stakeholders to align with each other as they come onboard. Furthermore, we hope to build on and refine these emerging lessons with the help of interactions between all the existing and new stakeholders. While these lessons emerged in the EdTech research-practice partnership, we believe that these will hold true for other RPPs in India as well which share similar stakeholder characteristics. Finally, given the goal of establishing a healthy EdTech ecosystem in India, we believe that these lessons will initiate discussions amongst the policymakers and national education strategists on how to sustain and institutionalize this ongoing work. Towards this long-term goal, the above co-design episodes and lessons demonstrate the importance of having multiple stakeholders aligned around one shared aim who are generating and sharing knowledge with each other. We hope that this work will inspire innovative ideas driven by policymakers and strategists for sustaining such RPPs where the voices of all stakeholders are represented.

Acknowledgments

We acknowledge the collective effort of the EdTech Tulna team from IIT Bombay, the Central Square Foundation, and the International Innovation Corps team working with the state government officials in India.

References

Coburn, C.E. & Penuel, W.R. (2016). Research-practice partnerships: Outcomes, dynamics, and open questions. *Educational Researcher, 45* (1), 48-54.

Coburn, C.E., Penuel, W.R., & Geil, K. (2013). Research-practice partnerships at the district level: A new strategy for leveraging research for educational improvement. *William T. Grant Foundation.*

Central Square Foundation - CSF (2021). www.centralsquarefoundation.org, retrieved 25 May 2021.

CSF EdTech Lab 1 Report (2019). Retrieved on 25 May, 2021 from https://centralsquarefoundation.org/wp-content/uploads/EdTech%20Lab%20Report_November%202019.pdf

EdReports. (2021). Retrieved 25 May 2021, from https://www.edreports.org/

Farrell, C. C., Harrison, C., & Coburn, C. E. (2019). "What the hell is this, and who the hell are you?" Role and identity negotiation in research-practice partnerships. AERA Open, 5(2), 2332858419849595.

Ferreyra, M., & Liang, P. (2012). Information asymmetry and equilibrium monitoring in education. *Journal of Public Economics, 96*, 237–254.

Fishman, B.J., Penuel, W.R., Allen, A.-R., Cheng, B.H., Sabelli, N. (2013). Design-based imple- mentation research: An emerging model for transforming the relationship of research and practice. In B. J. Fishman, W. R. Penuel, A.-R. Allen, & B. H. Cheng (Eds.), *Design-based implementation research. National Society for the Study of Education Yearbook*, *112*(2), 136–156.

Kay R. & Knaack L. (2009): Assessing Learning, Quality, and Engagement in Learning Objects: The Learning Object Evaluation Scale for Students (LOES-S). Edu. Tech. Research and Development, 57(2), 147-168

Leacock, T. L. & Nesbit, J. C. (2007): A framework for evaluating the quality of multimedia learning resources. Journal of Educational Technology and Society, 10(2), 44.

National Education Policy (2020). https://www.mhrd.gov.in/sites/upload_files/mhrd/files/nep/NEP_Final_English.pdf

NMEICT(2021). Retrieved May 31, 2021 from https://nmeict.ac.in/technology-enabled-learning/national-mission-in-education-through-ict/

Omidyar Network. (2019). Scaling access and impact: Realizing the power of EdTech. Omidyar Network. https://omidyar.com/scaling-access-impact-realizing-the-power-of-edtech/

Peng, H., Ma, S., & Spector, J. (2019). Personalized Adaptive Learning: An Emerging Pedagogical Approach Enabled by a Smart Learning Environment. In *Lecture Notes in Educational Technology* (pp. 171–176).

Penuel, W., Allen, A.-R., Coburn, C., & Farrell, C. (2015). Conceptualizing Research–Practice Partnerships as Joint Work at Boundaries. *Journal of Education for Students Placed at Risk (JESPAR)*, *20*, 182–197.

Penuel, W. (2019). Co-design as Infrastructuring with Attention to Power: Building Collective Capacity for Equitable Teaching and Learning Through Design-Based Implementation Research (pp. 387–401).

Quintana, C., Reiser, B., Davis, E., Krajcik, J., Fretz, E., Duncan, R., … Soloway, E. (2004). A Scaffolding Design Framework for Software to Support Science Inquiry. *J. of the Learning Sciences*, *13*, 337–386.

Severance, S., Leary, H., & Johnson, R. (2014). Tensions in a Multi-Tiered Research-Practice Partnership. In *Proceedings of International Conference of the Learning Sciences, ICLS* (Vol. 2).

Tatar, D. (2007). The Design tensions framework. Human-Computer Interaction, 22, 413-451

Tseng, V. (2012). Partnerships: Shifting the dynamics between research and practice. New York, NY: William T. Grant Foundation.

Tseng, V., Easton, J., & Supplee, L. (2017). Research-Practice Partnerships: Building Two-Way Streets of Engagement. *Social Policy Report, 30*, 1–17.

Tulna (2021). https://edtechtulna.org/, retrieved 25 May 2021.

Alternative Approach for Evaluation Adapted for Times of Emergent Conditions

Dan KOHEN-VACS* & Meital AMZALAG
Holon Institute of Technology (HIT), Israel
*mrkohen@hit.ac.il

Abstract: In this paper, we propose a unique approach for an alternative evaluation, designed, deployed, and optimized for times of crises such as those experienced during the eruption of the COVID-19 pandemic. In our approach, we rely on technological tools for supporting interactions conducted among lecturers and students. We suggest this evaluation activity incorporated with four phases that are interrelated. Thus, technologically supported interactions from an early phase are used in a later phase. This is done in a manner emphasizing the interdependency and chronological order of the phases practiced along with the evaluation. Our design efforts were exercised as we considered and aligned it to four meta-principles suggested by other researchers and widely adopted by researchers educators from around the world. Here we demonstrate the conduction of our research efforts as we deploy the evaluation approach in programming courses corresponding to a bachelor level. Our exploration efforts were exercised as we relied on qualitative methods used for analyzing quotes mentioned by 18 students attending semi-structured interviews. The findings of this research indicate that students perceive this evaluation activity as effective. Specifically, these students mentioned that this approach encouraged their capabilities to be autonomous in their learning process. Additionally, it fostered their creativity and emphasized their perception of the importance of a lifelong learning approach. Based on the mentioned principles and meta-principles we addressed in this paper, we suggest the existence of a new level of meta-principles consolidating the four ones focused on our efforts. There we address the current leaner expected to become a future citizen in a dynamic, flexible and technological society possibly encountering other challenges beyond the ones tackled in times of COVID-19. Accordingly, our suggestion for an alternative approach for evaluation is proposed as an educational pattern usable for the time of future normality as well as a way to cope with educational challenges in moments of future crises.

Keywords: Alternative evaluation approach, programming courses, higher education, meta-principles, COVID-19, emergent conditions

1. Introduction

In recent decades, organizations practicing education are increasingly impacted by changes resulting from various and possible factors, including social aspects, economic-related shifts, and other emergent crises potentially influencing their professionals and students. In such conditions, they must address such situations by adapting their efforts in manners capable of coping with such challenges(Chaturvedi et al., 2021; Fontdevila et al., 2021; Levchenko et al., 2017). To cope with crises consisting of dynamic conditions, educational organizations may select to employ innovative approaches exercised as part of flexible and innovative thinking. Moreover, in many cases, they do so while still relying on professional principles adopted and adapted to the context of emergent challenges (Dhawan 2020; Marouli, 2021). In this respect, kali and other searchers established a growing Design Principles Database (DPD) containing tools offered to educational practitioners and researchers as a set of convenient means for authoring their educational (Kali, 2006; Kidron & Kali, 2017). There, they suggest principles usable as design tools representing also exploitable as compass assisting lectures in their practices (Sagy et al., 2018). This database includes principles relying on various theoretical frameworks such as constructivism, socio-constructivism, constructionism (McKenney & Reeves, 2018).

The DPD also includes high and comprehensive principles addressed as meta-principles. For example, one meta-principle addresses (a) making student's thinking visible. This meta-principle offers benefits for group-learning considering to practice efforts encompassed by documentation of the educational process. This meta-principle could be beneficial both for individual or group-work aspiring to reflect their educational process and outcomes beyond the boundaries of the classroom (Richardson et al., 2018). Another meta-principle addresses the (b) making of science accessible reflects repositioning the lecturer's role as the main source of knowledge. Thus, converting traditional teaching style to a more coaching type of education, fostering fruitful discussions enabling and students with better outcomes resulted from their individuals or group-work (Kizilaslan et al., 2019). In addition and in light of group work, the mentioned meta-principles, also emphasized collaborative learning in the sense of (c) helping students to learn from each other (Gillies & R.M., 2019). This meta-principle reflects the fostering dissemination of knowledge among peers involved in a collaborative learning process, potentially benefitting individual students as well as groups. Last but not least, we present another meta-principle addressing the (d) promotion of autonomous learning. This meta-principle reflects students encouraged to take best decisions and eventually capable to become self-reliant that may keep exercising their efforts also outside of the framework of the official classroom (Andriani et al., 2018). As implied, there is a growing community of researchers seeking to propose additional principles possible exploited and explored by educational stakeholders searching for additional means and approaches to design innovative, meaningful, relevant, and appealing activities representing state-of-the-art education (Kali et al., 2020; Linn et al., 2018). Moreover, such intends to refine and excel design practices of educational activities are demanded as a result of ongoing shifts in communities concerning various changes it may gradually experience. Additionally, such acknowledgment of principles may become crucial in cases of communities tackled by emergent conditions, suddenly demanded to make instant adaptations in their learning practices as a result of crises it is experiencing.

For the past year, educational institutions located worldwide, are facing the consequences of such a crisis known as COVID-19 (Elman et al, 2020). This pandemic is influencing various aspects of educational routines lecturers and students used to exercise on daily basis. Many of these consequences are emphasized in form of social isolation imposed and still impacting many schools and universities. In some extreme cases, schools shifted their mode of operation to remote learning. Accordingly and as part of intensions of education institutions struggling to cope and overcome educational challenges related to social isolation, they exploited their existing Learning Management Systems (LMS). Occasionally, these type of systems is used along with other tools such as google classroom or zoom enabling the transition of the traditional and synchronous classroom to the mentioned and emergent conditions (Cornock, 2020; Reynolds & Chu, 2020). In this respect, many of these educational organizations quickly discovered that technical transition should address design requirements concerning educational experiences adapted for such challenging times (König et al., 2020). This requirement for adaptation also concerns activities focusing on evaluation experiences practiced as part of comprehensive and educational activities challenges (Karalis, 2020).Accordingly and in this respects, in this paper, we present our suggestion for an alternative evaluation approach adapted for the mentioned circumstances. Specifically, an alternative approach for assessing students taking programming courses in bachelor level attending in higher education in times of social isolation imposing remote-learning. In this respect, we emphasize that we initiated our efforts as traditional exams were not possible to conduct (in many places still not possible).

Here, we present our efforts, to design, deploy and explore our evaluation approach encompassed by technological support. Specifically, we deploy a multiphase activity requiring lecturers and students to use technology for participating in a workshop practiced for evaluation purposes. As implied, we design this activity while seeking to foster well-known values and principles serving as a compass to orient practitioners' efforts to offer an activity representing state-of-the-art practice. We emphasize that this activity should be exercised while practicing innovative and flexible thinking. We deploy this activity as part of our aspiring suggest an evaluation activity that is technologically supported capable to offer a convenient way to cope with requirements for an evaluation in educational settings in light of emergent conditions such as those imposed on lecturers and students in times of COVID-19.

2. Research Aim and Research Question

In this research effort, we focus on two main aims. The 1st one deals with the examination of our deployment efforts. Thus, examining the adoption of previous practices implemented in traditional evaluations, adopted and consolidated into a unique evaluation approach optimized for coping with emergent challenges tackled in times of COVID-19. In this aim, we focus our observation on how the proceeding of this evaluation as a multiphase activity works in the realistic settings of the classroom conducted in those challenging moments. Additionally and in a further research aim, we examine the alignment of our evaluation approach to meta-principles mentioned in the previous section. As implied, one of the meta-principles required some adjustment in order to adapt it to the setting for which we suggest our evaluation activity (programming courses). In this respect, we examine the evaluation in light of the mentioned meta principles as we suggest an adjustment of the 2nd one, reformulating it to the repositioning of the lecturer's role in light of the programming courses aspired to made the corresponded content knowledge visible.

We postulate these research aims as we have in mind that along with our current efforts, we might discover additional and new principles also aligned to the proceeding of the practiced activity. Accordingly, we suggest the following research question:

How and to what extend the four meta-principles (as described in the introduction), manifest themselves along with the proceeding of the evaluation activity deployed in times of COVID-19?

3. Methodological Approach

In this research effort, we practiced a qualitative approach as we conducted Zoom interviews with students participating in the mentioned evaluation. The average duration of the interviews was half an hour. Interviewees were asked about their feeling concerning the affordance to select their own challenge and theme for their evaluation. In addition and during the interviewees, we examined their positions concerning the contribution of the formative learning and feedback that experience during the learning process. They were also required to address the benefits and challenges concerning the alternative evaluation method. Last but not least and as normally practiced in such interviews, we left some time allowing interviewees to bring additional comments addressing the evaluation that they consider interesting.

We then conducted a content analysis based on work exercised by researchers, each separately ascribing the transcripts of interviews according to topics and themes. Next, researchers compared the outcomes of their work. Since that compared aspects resulted in unsatisfactory similarities in terms of corresponded topics and themes, researchers discussed, clarified, and agreed on each of the quotations as they corresponded to an agreed category. Researchers proceed to data analysis only after achieving 80% compatibility of their categorization efforts. This research was conducted as part of two semestrial courses: Introduction to Programing, during the summer semester of 2020 (first-year students), and Programming-1 during spring semester 2020 (second-year students).

3.1 Evaluation Design

In this section, we present the structure of the evaluation activity including its four phases, along with the technologies we exploit to support each of them. The evaluation activity consisted of several phases which are interdependent on each other. All of the phases require lecturers and students to use technologies for various purposes including the Zoom tool exploited for conduction of synchronous meetings and Moodle LMS used for accessing the content in form of instructions needed for the evaluation and course content permitted to be used during the evaluation. Last but not least, occasionally, students could use Google Docs in order to share the editable documents (in real-time) as they formulate their own challenge corresponding to a theme from their realistic settings. In Figure 1, we demonstrate the evaluation activity, its flow of phases, and the technology used for supporting communication in each and across them.

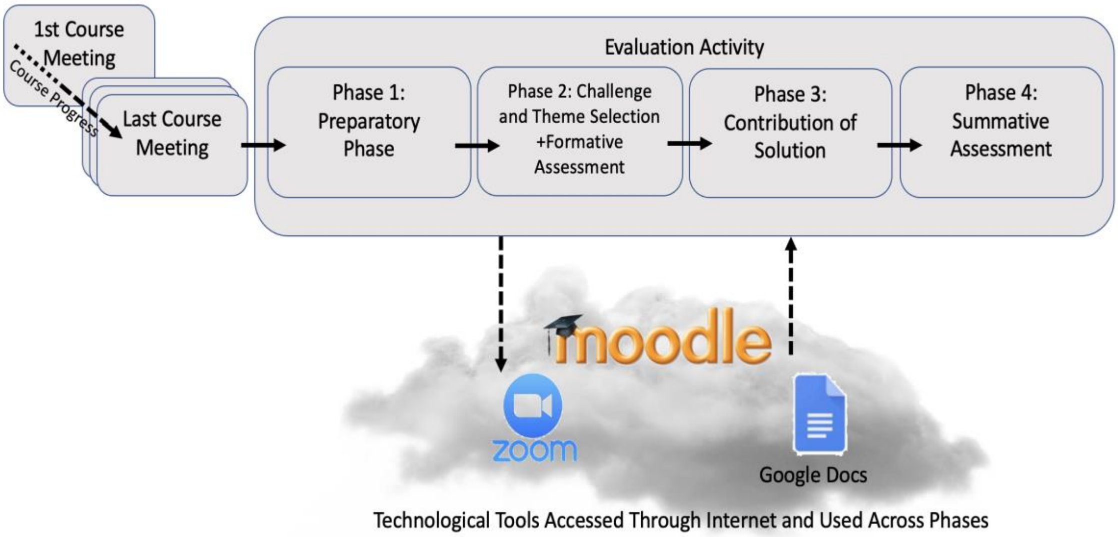

1st Course Meeting
Course Progress
Last Course Meeting
Evaluation Activity
Phase 1: Preparatory Phase
Phase 2: Challenge and Theme Selection +Formative Assessment
Phase 3: Contribution of Solution
Phase 4: Summative Assessment
moodle
zoom
Google Docs
Technological Tools Accessed Through Internet and Used Across Phases

Figure 1. Illustration of Activity including its Interactive Phases which are Technologically Supported.

The figure above describes the activity spanning along with four phases of interactions practiced between lectures and students. In the 1st phase, students are being prepared for this activity. This phase is used in order to serve as a means for making steps between the course meetings and the evaluation activity more seamless. We do so as we consider this evaluation as a comprehensive part of the evaluation activity conducted following the course. Accordingly, we would like to communicate with students with this notion, making them feel that this is another step along the programming course and not a test that occasionally is perceived as a differential from the course. In this phase, we used Zoom for conducting online meetings. We also used Moodle LMS for accessing instructions for the evaluation as well as learning materials from the course.

In the next phase, students were instructed to formulate their own challenges they selected for evaluation. In this respect, they were instructed to formulate a question and were given the option to suggest a challenge reflecting a theme (topic or situation) from their realistic settings. We provided such affordance as we considered these aspects with potentials for enhancing students' engagement to the process exercised along with the evaluation. Later and in the same phase, lecturers and students meet on 1:1 synchronous meeting. This meeting was conducted in order to verify and adjust the challenge in an agreed way verifying students' suggestions includes meaningful content properly addressing the content learned during the course. Additionally, this was conducted in order to ensure that students' formulation of their own challenges are meaningful in light of the aims of the evaluation. In this phase, we used technologies including Zoom for conducting online meetings. We also used Moodle LMS for accessing instructions for the evaluation as well as learning materials from the course. Occasionally in this phase, we offered students to use Google Forms enabling common editing of a document usable for collaboratively formulation of students' challenge. Thus, for ensuring in a synchronous meeting between the lecturer and the student that challenge is appropriate for the aims of the evaluation.

The 3rd phase includes students' solutions of their own and adjusted challenges. Here they could summit lecturers to zoom meetings for asking different questions concerning clarification of the instructions of the evaluation (in the same way that is done in a traditional test). For this phase, we used Moodle LMS for accessing instructions for the evaluation as well as learning materials from the course.

The 4th and last phase is conducted following the submission of the authored challenge and solution as a PDF and code file submitted to a Moodle LMS. Then, the student attends a short (5 to 10 minutes) defense session in which he is being asked about different aspects of his own solution. In this sense, students are evaluated on the solution and not on the question he/she authored. In this phase, we mainly used Zoom for conducting the synchronous meeting.

3.2 *Participants*

The Interviews were conducted with 18 students attending bachelor degree in the domain of social sciences: 13 female and 5 male who took the course on Introduction to Programing, and 4 female and one male who attended the programming 1 course. All the mentioned students voluntarily agreed to participate in semi-structured interviews focused on their experiences prior, during, and following they took the evaluation activity.

3.3 Ethical Considerations

In respect to ethical considerations, we emphasize that an authorization to conduct this research was approved from the comity for ethical affairs of the institution to which the researchers are affiliated. As part of this authorization and since one of the researchers was the lecturer in the mentioned courses, the authority instructed the researchers to exercise these efforts while emphasizing that the 2nd researcher will conduct the semi-structured interviews. Interviewees were told that they are allowed and for whatever reason to stop their participation in the interview (without jeopardizing them in any way). The researcher that conducted the interviews was responsible for making anonymized transcriptions. Thus, the 1st researcher was not able to correspond a comment with a specific student taking the interview. Interviews were conducted only following grade publishing in the courses. Conclusively, the research was authorized under the mentioned conditions and researchers accepted the conditions and conducted their efforts accordingly.

4. Findings

As mentioned and in the methodological section, we exercise our research efforts as we qualitatively analyze, data collected during semi-structured interviews. In this section, we bring a sample of collected data reflecting transcriptions of interviews conducted with 18 students taking both programming courses. The quotations brought in this section are presented according to the four phases of the evaluation and in correspondence of a meta-principle, they reflect.

4.1 Phase 1: Preparatory Phase

In this part, we address the preparatory phase referring all to the sessions prior to the evaluation experience. Student O. commented in respect to the meta-principle focussed on helping students to learn from each other and mentioned that: *“lecturers prepared us very well and provided us with a precipice explanation concerning the phases to be conducted along with the evaluation. I was not surprised and this is really great…Additionally, lecturers provided us with a stimulatory session to the assessment prior to the real experience.*

As for the meta-principle concerning the promotion of autonomous learning as well as another one concerning the repositions of the lecturer’s role, student I2 told that:” *As the lessons in this course concluded, I was encouraged to gather and revisit all learning material and assignments. I even formulated various versions and combinations of new exercises based on the assignment given during the course. It made me feel totally responsible to my learning process*”. In addition, student Y2 told that: “*We concluded the lessons of the course, and I was encouraged to review the materials we studied along the course. I even formulated self exercises mixing and combining different topics we covered in the Couse. It provided me the opportunity to experience formulation of my own challenges oriented to cover all materials we ever learned in the course. Just from these rehearsals, I completely sure that I am ready for any type of evaluation. I felt responsible for the implementation to the knowledge I acquired along the course* ”. student L. said that this type of preparation fostered a set of mind which is more open, flexible and therefore made him more ready for the test.

From the mentioned quotations, we observe that students perceived themselves as responsible for their own learning process. Additionally, they expressed satisfaction as they manage to solve new challenges consolidating aspects focused along the course. They did it as they formulate their own questions consolidating their own selected challenge presented in light of a theme taken from their realistic setting. Additionally, their efforts reflect a meaningful learning process as they cope with challenges, intended to be solved along a process consisting of coding, debugging, and correction in phases of evaluation. Last but not least, students’ capabilities to act autonomously was emphasized. In this sense, we notice that they did not mention the lecturer in respect of the solution process. In this light and accordingly, the quotations illustrate three meta-principles reflected during the preparatory phase. We emphasize that the fourth principle concerning making learning visible was not identified in students' quotations corresponding to the preparatory phase. Additionally, we assume that the meta-principle concerning helping students to learn from each other would not be manifested in further phases of the evaluation. This is expected to happen as the students we instructed to take the evaluation as individuals.

4.2 Phase 2: Challenge and Theme Selection +Formative Assessment

In this phase, students were requested to select a challenge and a theme for the question they formulate for themselves. Students were encouraged to select a topic of the challenge and the theme that they may sympathize with possibly taken from their real life. Student T. mentioned that: "*I choose theme concerning real-estate companies and it helped me a lot as I used to work in such business and it represents a familiar domain for me*".

In the following quotations, we bring sentences indicating that the formulation of their own challenge implies a process encouraging students to practice higher thinking levels. Student T. mentioned that: "*personally, I think that the self-formulation of a challenge reflects an autonomous and higher thinking level. It reflects that such evaluation demands a higher level of mastery compared to traditional tests*". Another student named Sh. Told that: "*I think that when you knowing how to ask is much more meaningful than knowing the answer to a question*". Student R. added that "*formulating my own challenge represent a big advantage to me. I selected a Pizzeria as a theme for my question and such topic I sympathize assisted me while thinking in my own schemes*". Student L. told us that: "*Learning something while asking the question is a better approach offering support and encouraging students' creativity*". He continued and mentioned that: "*Formulating my own questions or challenges is helpful and this is something I take with me from this test experience. I think that this educational approach is applicable across subject matter as I practiced the same approach in a course on psychology as well as in another course on Research-Methods. Thus in all these courses I asked myself questions and developed my own thinking skills*".

Students' quotations indicate the cruciality of the affordance providing them opportunities to formulate their own challenge and use a theme from their realistic settings. Moreover, in such a way students' engagement in the evaluation experience is enhanced and represents another of various benefits of the evaluation activity. These benefits reflect various meta-principles including promotion of autonomous student, making thinking visible, and repositioning of lecturer's role. Additionally and in respect to some quotations, students implied on engorgement to adopt lifelong learning approaches possibly representing an additional principle beyond the mentioned four meta-principles.

4.3 Phase 3: Contribution of Solution

In this phase, students were supposed to answer the question they formulated and then adjusted by lecturer. Student P. said that: "*The exam in this format permitted me to better express myself as I had the opportunity to show there my acquired skills*". Student G. pointed out that during the evaluation he used presentations and examples provided along the course. He added that: "*I felt that I am concluding the evaluation with more knowledge compared to what I felt that I know prior to that experience*". Student Y. mentioned that "*as I began to solve my own challenge, I felt that something did not work well…It caused me to reconsider and decompose my challenge into small units each corresponding to a different topic learned during the course*". Student Sh. Concluded and pointed that the fact that "we solved our challenges on the real development environment enabled us to rethink things, debug and correct….all resulting in a much more meaningful learning process".

Student Y2., mentioned that: "*the evaluation provided me with a great opportunity to experience coding in real-life conditions. I acknowledge that occasionally the lecturers won't have answers to every question and therefore I'm required to a kind of autodidacticism skils…I have no dough that this approach represents an excellent and simple way to learn while being evaluated while understanding things instead of memorizing*".

In the mentioned quotations, students pointed out that they concluded the evaluation activity with more knowledge compared to their perceived amount of knowledge prior to their engagement to this experience. Students mentioned real-time debugging practiced during the proceeding of the evaluation reflecting their capabilities to act as autonomous students. Accordingly, we consider that these quotations reflect various meta-principles including promotion of autonomous learning, repositioning of lecturer’s role, and making thinking visible. In addition and following some of the quotations, we observe that students are concerned and address the demands they are expected to face in the professional world they would meet in the modern and labor market.

4.4 Phase 4: Summative Assessment

The last phase was conducted following to submission of students’ answers to their own formulated challenge that was later adjusted by the lectures in the course. Student I. mentioned that “*I had to attend the defense following the submission of my solution, I arrived at this phase while being quite confident with my solution and therefore it felt easier to answer my own questions*”. Student Sh. added that: “*During the defense, as I built my own solution, it felt easy and I wasn’t nervous or stressed*”. Another student named L. said that “*When I entered to my defense, I calmly talked to the lectures and explained to them the elements of my code and I showed to them my understanding in it*”. Student Y2., commented that during this phase, teachers asked him to explain the program he coded and also required him to provide some elaboration on the thinking process exercised during the coding. Last but not least, student P. told that: “I precisely understood for which aspects I would gain a lower grade. Thus, I understood exactly my mistakes!”.

In these quotations, students emphasized their formulation of the challenge as its corresponding solution and therefore they approach the evaluation calmly and confidently. In this light, we point out that quotations reflect various meta-principles including promotion of autonomous learning and making thinking visible.

In this paragraph, we conclude and overall address all the phases as described in previous subsections. The mentioned quotation indicates over benefits of our proposed approach for evaluation approach. Here, we acknowledge that all four Meta principles outlined and are clearly manifested (along with quotations) in respect of all the four phases of the evaluation activity.

5. Discussion and Concluding Remarks

In this paper, we suggest a unique approach for a workshop activity serving evaluation purposes usable in programming courses taken by bachelor students. We deployed an activity consisting of four and interrelated phases in which lecturers and students used technological means for interacting with each other. This unique experience is suggested in response to the emergent conditions as presented in times of the COVID-19 pandemic. In this sense, the exploitation of technological means is crucial in light of conditions consisting of social isolation. Thus the design of interactions exercised along the phases was planned to be conducted in synchronous as well as asynchronous modes using technological tools for communication (Zoom), managing learning content (Moodle) and collaborative formulation of documents usable by lecturers and students collaboratively adjusting challenges to be coped during the evaluation activity (Google Docs). In this sense, we notice that the use of technological tools felt seamless and “transparent” for both lecturers and students converting these aspects to a non-issue.

We designed the evaluation activity as we considered four meta-principles as proposed by Kali and other researchers (Kali, 2006; Sagy et al., 2018). Specifically, one of the principles required a mild adjustment fitting it to the context of programming courses. In this sense, the aspect of making science visible was addressed here as making the course content (from programming courses) visible. In the quotations collected from the semi-structured interviews conducted with students attending the programming courses, we found that all four meta-principles are manifested along with the phases of the evaluation activity. The findings in these research efforts also indicate possible principles applicable to the proceedings of the evaluation activity. We found that the principle concerning lifelong learning is applicable to the aims of the evaluation and as reflected from students' quotations. We consider this principle critical for the development of citizenship in light of advanced, dynamic and technological societies. Additionally, this principle becomes crucial for fostering intellectual development and professional skills. Last but not least, it is also important for improving one’s quality of life (Boeren & Whittaker, 2018; Tatarinceva et al., 2018). Moreover, this principle is becoming critical in light of the dynamics of daily routines in the 21st century (Tatarinceva et al., 2018). We acknowledge that in order to encourage learners to adopt a lifelong learning approach, they need to be communicated and convinced on the relevance of content to be focused along their learning path (Bråten et al., 2018).

Accordingly, we propose that lifelong learning, as a principle dependent on relevancy, is crucial for optimizing the learning process as a means adapted for civic life and the labor market of the 21st century. Thus, we suggest the possibility to use higher level and additional meta-principle relying on the four meta-principles mentioned in this paper (Kali, 2006). We do so as we acknowledge that civic life and the labor market in the 21st century will present future citizens with a dynamic, flexible, and technological environment requiring peoples' creativity and higher thinking skills. Additionally, we suggest that current students (and future citizens), should be capable to access and exploit data and information from various and decentralized sources. They need to be able to work collaboratively in various group settings. Furthermore, they need to adopt capabilities making them autonomous learners. In order to capacitate current students (future workers) capable to successfully cope with the demands of the 21st-century labor market, we recommend the consideration and possible adaptations of all the mentioned properties. Furthermore, and while considering the cruciality of these sets of demands, we suggest the capacitation for the 21st-century market as a higher meta-principles relying on the four ones we focused on in this paper (Kali, 2006).

We consider our evaluation activity as an approach capable to address challenges corresponding in times of normality as well as of challenging moments such as those encountered in times of covid-19. In this respect and in correspondence to the mentioned in this section, we consider normality in the sense of modern life in the 21st century including requirements that would be demanded from citizens in this epoch. As mentioned we also aim this approach for challenging moments. Therefore, we emphasize that we offer our evaluation approach as an activity usable in other moments of crisis consisting of new types of emergent situations.

6. Future Efforts

In our future efforts, we intend to continue our design, development, and deployment activities focused on the refinement of our unique evaluation approach for additional subject domains. In this sense, we intend to explore the applicability of this approach across disciplines. Thus, we aim to examine and accordingly refine this approach beyond the domain of programming courses. Additionally, we acknowledge that the deployment of such an evaluation approach might be experienced differently depending on the organizational nature of the educational instruction practicing it. Therefore, in our future efforts, we intend to deploy the evaluation across institutions and by different lecturers from across domains as well as from different cultural environments corresponding to their institutions. We also aim to continue and explore additional principles and meta-principles applicable to our suggested approach for evaluation. Last and not least, we acknowledge that COVID-19 may represent a single example of a crisis among others humanity may tackle in the future. Therefore, we aim to suggest, explore and refine our efforts while suggesting this activity as an educational pattern that could be adopted and adapted for normal routines as well as for emergent conditions societies may encounter in future crises. In The coming efforts we intend to focus on the mentioned directions for research while using mixed-method of methodological approaches.

7. Limitations of Research

In this section, we bring several limitations concerning our research efforts. We acknowledge that this research was conducted with a relatively small amount of interviewees. Additionally, we used here a qualitative method of research. Finally, we focused our research efforts on two courses conducted in a single institution dealing with the same subject domain.

References

Andriani, P. F., Padmadewi, N. N., & Budasi, I. G. (2018). Promoting autonomous learning in English through the implementation of Content and Language Integrated Learning (CLIL) in *science and maths subjects*. In SHS Web of Conferences (Vol. 42, p. 00074). EDP Sciences.

Boeren, E., & Whittaker, S. (2018). A typology of education and training provisions for low educated adults: categories and definitions. *Studies in the Education of Adults*, *50*(1), 4-18.

Bråten, I., McCrudden, M. T., Stang Lund, E., Brante, E. W., & Strømsø, H. I. (2018). Task‐oriented learning with multiple documents: Effects of topic familiarity, author expertise, and content relevance on document selection, processing, and use. *Reading Research Quarterly, 53*(3), 345-365.

Chaturvedi, K., Vishwakarma, D. K., & Singh, N. (2021). COVID-19 and its impact on education, social life and mental health of students: A survey. *Children and youth services review*, 121, 105866.

Cornock, M. (2020). Scaling up online learning during the Covid-19virus (Covid-19) pandemic. Retrieved from: https://mattcornock.co.uk/technology-enhanced-learning/scaling-up-online-learning-during-the-Covid-19virus-covid-19-pandemic/

Dhawan, S. (2020). Online learning: A panacea in the time of COVID-19 crisis. *Journal of Educational Technology Systems*, *49*(1), 5-22.

Elman, A., Breckman, R., Clark, S., Gottesman, E., Rachmuth, L., Reiff, M. Et al. (2020). Effects of the Covid-19 outbreak on elder mistreatment and response in New York City: Initial lessons. *Journal of Applied Gerontology*, *39*(7), 690-699.

Fontdevila, C., Verger, A., & Avelar, M. (2021). The business of policy: a review of the corporate sector's emerging strategies in the promotion of education reform. *Critical studies in education*, *62*(2), 131-146.

Gillies, R. M. (2019). Promoting academically productive student dialogue during collaborative learning. *International Journal of Educational Research*, 97, 200-209.

Kali, Y. (2006). Collaborative knowledge building using the Design Principles Database. International Journal of *Computer-Supported Collaborative Learning*, *1*(2), 187-201.

Kali, Y., Sagy, O., Lavie-Alon, N., Dolev, R., & Center, T. C. S. S. (2020). From a Network of Research-Practice Partnerships to a multi-expertise learning and design community. 10.13140/RG.2.2.22169.24163.

Karalis, T. (2020). Planning and evaluation during educational disruption: Lessons student from COVID-19 pandemic for treatment of emergencies in education. *European Journal of Education Studies*.

König, J., Jäger-Biela, D. J., & Glutsch, N. (2020). Adapting to online teaching during COVID-19 school closure: teacher education and teacher competence effects among early career teachers in Germany. *European Journal of Teacher Education*, *43*(4), 608-622.

Kidron, A., & Kali, Y. (2017). Extending the applicability of design-based research through research-practice partnerships. EDeR. *Educational Design Research, 1*(2).

Kizilaslan, A., Sozbilir, M., & Zorluoglu, S. L. (2019). Making Science Accessible to Students with Visual Impairments: Insulation-Materials Investigation. *Journal of Chemical Education*, *96*(7), 1383-1388.

Levchenko, O., Levchenko, A., Horpynchenko, O., & Tsarenko, I. (2017). The impact of higher education on national economic and social development: comparative analysis. *Journal of Applied Economic Sciences*, *3*(49): 850–62.

Linn, M. C., McElhaney, K. W., Gerard, L., & Matuk, C. (2018). Inquiry learning and opportunities for technology. In *International handbook of the learning sciences* (pp. 221-233). Routledge.

Marouli, C. (2021). Sustainability Education for the Future? Challenges and Implications for Education and Pedagogy in the 21st Century. *Sustainability*, *13*(5), 2901.

McKenney, S., & Reeves, T. C. (2018). Conducting educational design research. Routledge. doi:10.4324/9780203818183.

Reynolds, R., & Chu, S.K.W. (2020), Guest editorial. *Information and Learning Sciences*, *121*(5/6), 233-239. Retrieved from: https://doi.org/10.1108/ILS-05-2020-144

Richardson, G. M., Byrne, L. L., & Liang, L. L. (2018). Making learning visible: Developing preservice teachers' pedagogical content knowledge and teaching efficacy beliefs in environmental education. *Applied Environmental Education & Communication*, *17*(1), 41-56.

Sagy, O., Kali, Y., Tsaushu, M., & Tal, T. (2018). The Culture of Learning Continuum: promoting internal values in higher education. *Studies in Higher Education*, *43*(3), 416-436.

Tatarinceva, A. M., Sokolova, N. L., Mrachenko, E. A., Sergeeva, M. G., & Samokhin, I. S. (2018). Factors determining individual success in Life-long learning. Espacios, 39(2), 29.

Wiranegara, D. A., & Hairi, S. (2020). Conducting English learning activities by implementing Telegram group class during COVID-19 pandemic. *Journal of English for Academic and Specific Purposes*, *3*(2), 104-114.

Cura Personalis: Institutionalizing Compassion during Emergency Remote Teaching

Ma. Monica L. MORENO*, Maria Mercedes T. RODRIGO, Johanna Marion R. TORRES, Timothy Jireh GASPAR & Jenilyn L. AGAPITO
Ateneo de Manila University, Philippines
*mmoreno@ateneo.edu

Abstract: Faced with the fears and anxieties brought on by the COVID-19 crisis, educational institutions had to devise new compassion-based teaching and learning policies and approaches that recognized and provided for the pandemic's psychological and emotional toll. This paper describes how the Ateneo de Manila University in the Philippines enacted its core value of *cura personalis*, care for the entire person, in the context of emergency remote teaching. We describe the circumstances that prompted the greater emphasis on compassion and the adjustments to classroom management, course content, class interactions, and assessment. Finally we describe the tradeoffs or costs of this approach.

Keywords: Compassion-based teaching, COVID-19, cura personalis, emergency remote teaching, teaching and learning practices, higher education

1. Introduction

The emergence of the COVID-19 virus in early 2020 forced schools, colleges and universities worldwide to convert face-to-face courses to Emergency Remote Teaching (ERT) to continue serving the educational needs of approximately 1.5 billion students worldwide (UNESCO, 2020). ERT refers to the temporary shift of instructional modality from a face-to-face or blended mode to an alternate mode such as mobile learning, radio or other methods that are contextually feasible. ERT must be distinguished from online learning, which is "a form of distance education in which a course or a program is intentionally designed in advance to be delivered fully online." (Bates, 2020). In online learning, teaching and learning methods are designed to maximize the affordances of the instructional mode. In contrast, ERT is not planned; rather, it is characterized by rapid response, adjustment, and accommodation to the extent possible. Research studies on ERT experiences of the higher education sector with COVID-19 have ranged from comparisons between national initiatives for academic continuity (Coutts et al., 2020), to the documentation of administrative and faculty responses to ERT (Johnson et al., 2020), the management of inequities and student needs (Baker, 2020), and short- and long-term implications for academic institutions (Veletsianos & Kimmons, 2020). These studies affirm that due to the massive disruption in higher education arising from COVID-19, ERT was a universal response to ensure academic continuity. In its simplest form, ERT meant powering through what remained of the curriculum. However, during that time, students were experiencing tremendous strain: some needed to vacate their on-campus accommodations on short notice; others had increased family or financial responsibilities; and many were distressed about their health and safety (Bozkurt et al, 2020). In the Philippines, many underprivileged learners did not have access to digital devices and the Internet, and Internet cafes were closed, making access to online classes difficult, if not impossible. Educational institutions had to revisit their priorities and devise new compassion-based policies that recognized and provided for the pandemic's psychological and emotional toll. Schools could not treat students as "cognitive machines" (Bozkurt et al., 2020). Instead, a pedagogy of care, one that was always needed and had always existed, had to become more visible and more palpable during the COVID-19 crisis.

The Ateneo De Manila University (ADMU) in the Philippines was no exception. It shifted to ERT from March 16 till May 8, 2020. While doing so, it enacted one of its core values, *cura personalis*, care for the entire person, to arrive at a pedagogy that prioritized compassion. Compassion is defined as the recognition of the suffering of others and the decision to take action to alleviate that suffering

(Gelles et al., 2020). It has four main components: an awareness of the suffering (cognitive), a sympathetic concern (affective), a desire to relieve the suffering (intentional), and a readiness to help relieve the suffering (motivational). Other studies regard compassion as a three-part process involving noticing another's pain, feeling for the other's pain, and acting in response to the pain (Frost et al., 2006). Based on this framework, institutional compassion is said to exist when members of a system collectively notice, feel and respond to pain experienced by members (Kanov et al., 2004). Focusing on institutional responses to suffering among its members allows us to understand how collective proactive, creative, and empathetic actions help bind organizations. In this paper, we attempt to describe how, as an institution, ADMU enacted compassion during its ERT. We answer three main questions: (1) What were the circumstances that motivated the emphasis on compassion? (2) What specific teaching and learning practices enacted this emphasis on compassion? (3) What were the tradeoffs or costs?

2. Methodology

The data reported in this paper was collected as part of a multi-institutional and multinational study that examined faculty, student, and administrative responses to ERT when COVID-19 first emerged in early 2020 (Bartolic et al., 2021). This paper's methods stem from that study. We focus on data collected from the Ateneo de Manila's Loyola Schools, a tertiary education institution that offers undergraduate and graduate degree programs. The Loyola Schools are headed by a Vice-President and divided into five schools headed by Deans: the School of Humanities, School of Social Sciences, School of Science and Engineering, the John Gokongwei School of Management and the Gokongwei Brothers School of Education and Learning Design. The Loyola Schools have approximately 10,000 students at the undergraduate and graduate levels.

2.1 Participants, Research Instruments, and Procedures

Five departments were pre-selected to represent samples of the university's disciplines: Computer Science, History, Psychology, Political Science, and Chemistry. Since the departments had unequal sizes, the Yamane formula was used to determine the percentage of courses and faculty that would form the sample of each department. Using the percentage generated by the formula, courses, names of faculty, and students were then randomly selected using MS Excel. The final number of faculty and students was 112 and 1032 respectively. From these, an actual total of 45 faculty members and 321 students agreed to participate in the study. Nineteen (19; 42%) faculty members had been teaching at ADMU for at least ten years; in contrast, twenty-six (26; 58%) had taught at the university for ten years or less. Twenty-seven faculty respondents (60%) had not taught an online course prior to the ERT period, while ten members (22%) stated they had previous experience with remote teaching. On the other hand, survey results were received from 108 first year students (34%), 123 second year students (38%), 56 third year students (17%) and thirty-four students in either their last year or in graduate study (11%). The common thread binding all participants was that each was part of an active course during the academic term in which the abrupt shift to ERT occurred.

Two data collection instruments were devised. The first was a virtual, semi-structured interview with faculty. This consisted of 70 core questions about one specific course that was taught during the ERT term. It included questions about course details before and after the transition to ERT in terms of classroom management, content delivery, interaction, and assessment. On average, the interviews lasted about 30 minutes to 1 hour per instructor. The second instrument was a web-based self-administered questionnaire on the same course which instructors completed prior to the interview. This sought their general feedback on the emergency transition to remote learning. A similar web-based self-administered questionnaire was given to student participants.

Each of the 112 course instructors and 1032 students was sent an email invitation to participate. Initial low student response rates resulted in the resending of the invitation to the entire student population of those taking courses offered by the aforesaid departments during the Second Semester of School Year 2019-2020. The invitation contained a statement about informed consent, as well as YES and NO buttons to indicate participation. A total of 45 faculty members and 321 students agreed to participate. This automatically recorded their email addresses and provided them a link to the self-administered questionnaire. It also notified the research team to send the faculty member an online interview invitation. Once scheduled, the researchers met the course instructor to record the virtual interview. All interviews were then transcribed; all but three interviews had been recorded. The responses from both the interviews and the student and faculty questionnaires were then summarized through frequency counts to answer the research questions. These were corroborated with data from faculty interviews and open-ended student questions. The researchers assumed that all faculty responses were honest and that the questions were answered to the best of the participants' ability.

2.2 Limitations

Due to the COVID-19 situation, data was collected using online forms and virtual interviews, which meant that the lack of any extralinguistic and non-verbal cues normally available in face-to-face settings may have affected the richness of the data (Hewson, 2015). Moreover, because of the relatively small sample of students, the results of the study cannot be generalized to the entire student population. Finally, the study aimed at retrieving general faculty and student experiences during the pivot to remote teaching; it does not compare experiences across year levels, disciplines and departments.

3. Results

Compassion on an institutional level can emerge when members of that institution first notice that pain is experienced by others (Kanov et al., 2004). This involves awareness of others' emotional states, and being attentive to emotional cues and experiences in one's context (Frost et al., 2006). The global pandemic meant that many individuals simultaneously felt anxious, uncertain, and devastated by the drastic changes imposed on society. Added to this was the abrupt shift to ERT spurred by the need for academic continuity, which caught both faculty and students off guard. The general feeling of anxiety and pain was observed on an institutional level. We now look at the physical, behavioral and psycho-emotional scenarios which prevailed during that period, prompting the need for compassion.

3.1 What were the Circumstances that Motivated the Emphasis on Compassion?

The interviews with the faculty shed light on some of the circumstances that underscored the need for compassion. The early days of the pandemic cultivated fear and panic. This resulted in a lack of focus and motivation to learn. When asked about their agreement with the statement "I was confident in my students' abilities to learn well in a remote online course", faculty responses were mixed. Twenty-three respondents (51%) agreed with the statement, but twelve (27%) did not. Ten respondents (22%) were neutral. This implies that while half of the faculty initially perceived their students to be capable of successfully transitioning to online learning, the other half did not share their confidence. During that time, the island of Luzon was placed under an Enhanced Community Quarantine (ECQ), shutting down travel, business, tourism, and face-to-face education. Faculty may have felt that the ECQ would also have affected students' mental and emotional states, possibly hindering them from a successful transition to online learning. One teacher shared that students "were worried about the pandemic [so] they lost focus, and that translated in their work."

The pandemic indeed gave students problems of their own. When asked about their agreement with the statement "As my instructor transitioned to remote online instruction, I was confident in my abilities to learn well in a remote online course", student responses were mixed. Half disagreed with the statement, 40% agreed, and the rest were neutral. This implied that many were not confident about their abilities to successfully hurdle the pivot to remote learning. Eighty percent felt overwhelmed by the transition, with half of the students believing it took more effort to complete work compared to before the transition. The majority (70%) of the students stated that the pivot to remote teaching resulted in a lower level of engagement ("not being motivated or engaged enough to do the tasks/readings"; "it was more difficult to find the motivation to do the tasks provided to us").

Fifty-two students (16%) stated they were concerned about the loss of motivation amidst the uncertainty and fear ("At that time, everything was hectic and uncertain so I did not feel that academics should be my top priority"). Another problem was technology (7%). Students needed a fast, reliable

Internet connection to participate in online classes, and this was not a universal resource. Beyond this, other student problems ranged from changes in living and working arrangements (14%) (“I do not have my own place to study, so often I am right beside the living room where my family watches TV and [goes] about their day”), to difficulties in self-learning without the benefit of real-time feedback (43%) (“most of the things we had to learn by ourselves”; “not being able to clarify something on the spot”). Students also had to confront a certain degree of loneliness. Without a set schedule and regular, contact with their peers, feelings of isolation started to grow (“it was harder to connect with classmates to collaborate and work together”). Table 1 shows the significant challenges mentioned by students.

Table 1. *Most Significant Challenges of Students during Remote Learning*

Challenges	n	Total (%)
Motivation	52	52 (16 %)
Internet connectivity	23	23 (7%)
Environmental factors	36	46 (14%)
School-life imbalance	10	
Lack of real-time feedback & guidance	68	137 (43%)
Keeping up with the online set-up	25	
Difficult to study on my own	23	
Difficulty in learning the material	21	
Lack of interaction with other students	32	32 (10%)

After the initial shock had worn off, feelings changed. Students had a realization that, “[it] looks like we’re going to be [in this situation] for a long, long time.” Students therefore started to ask how they would continue studying under the circumstances. In general, the transition to ERT affected students negatively. One teacher said, “Learning went down. I don't think they learned much after the transition and it's not their fault either.” However, there were students who continued to flourish. One teacher shared that, “I had some students who were highly driven…they were able to…channel the anxiety and use it for something more productive.” These students poured this energy into their work.

3.2 What Specific Teaching and Learning Practices enacted this Emphasis on Compassion?

After pain is noticed, the next two steps in the compassion process are feeling for the other’s pain and acting in response to the pain (Frost et al., 2006). On an institutional level, there is a collective feeling for others’ pain which involves not just empathizing with others, but going beyond feeling to involve a response to suffering to alleviate the others’ condition in some way. In this case, the collective response of the university was to enact compassionate flexibility (Gelles, 2020) in online teaching and learning practices to help students finish the courses as best as they could. These practices are discussed by teaching dimension: classroom management, content delivery, interaction, and assessment.

In terms of classroom management, one of the first things that teachers did was to establish ground rules for the ERT period. Forty-one instructors (91%) gave students written instructions on how to proceed online, and twenty-seven (60%) conducted an online interactive session with students to answer their questions about the. Eighteen faculty members (40%) posted a live streamed video or a YouTube-type video to explain how the course would be conducted remotely.

The ERT necessitated changes in content delivery. Many teachers (31/45, or 69%) defaulted to the use of live lectures conducted over the Internet during scheduled class hours. Three respondents stated they learned how to design more engaging content, and one instructor learned how to create podcasts. Thirty-one instructors (69%) also recorded videos of themselves. One teacher reported that, “I made instructional videos…and then just later uploaded [them] to support the asynchronous learning. [A] big part of my class relied on lab or hands-on activities, so I just uploaded instructional videos on YouTube for students to follow.” Some teachers made use of existing content with 35 respondents (78%) saying they made use of teaching materials they found online. The actual amount of content delivered had to be reduced drastically. This will be discussed in greater length in the next section.

Teachers supported class interaction with a range of online communication tools. The most popular were the use of broadcast email to the class (51%), and course announcements on an internet-based bulletin board or learning management system (LMS) (51%). Others conducted personalized individual communication using email, phone calls, Google Meet or Messenger. A fourth choice was the use of Facebook to post announcements, as well as the use of Discord, and text messaging. Teachers used a variety of avenues to maximize reach to students. On average, they communicated with their students 2-3 times a week. Teachers also took every opportunity to check in on them. These "*kamustahan*" ("how are you?") sessions became a regular part of class. One teacher would ask the class how the situation was affecting them ("...the students loved it because that gave them a semblance of normalcy at that time.") The "*kamustahan*" sessions gave the students and teachers a venue for open listening and empathizing with others, which facilitated the noticing and feeling aspects of compassion. More importantly, these provided a holding space for people to air their concerns and reflections, and where responses to aid healing could be given as the third aspect of compassion.

With regards to assessment, greater leniency was a recurring theme. Teachers did not give deductions for late submissions. Some modified the assignment or project specifications to better suit the online environment. Other teachers whose final projects required the citing of primary sources relaxed that requirement because students had no access to these texts. After three weeks in quarantine, it was becoming clear that both teachers and students were struggling. On 7 April 2020, the Vice-President for the Loyola Schools issued a memo declaring that all students would be given a passing grade for the semester. The memo stated that, "The Loyola Schools is of the mind to pass all eligible students this semester by giving them a P (pass) mark. Giving a P mark is the most humane way of dealing with student grades under the circumstances that we are in, where it is difficult and unfair to make a judgment of failure considering that students have not been given the benefit of a full semester to improve their performance."

3.3 What were the Tradeoffs or Costs?

The shift to ERT forced faculty to revisit their plans for the semester and trim content drastically. One teacher reported that, "I did take away academic topics, particularly…the last few modules that were more contemporary." Experiments that were usually conducted face-to-face had to be removed. Activities that involved oral presentations to a class also had to be withdrawn. One teacher noted that collaboration was difficult to achieve, especially among the freshmen. Discussing through Zoom or Facebook is not the same. "Half the battle for freshmen is getting to know their peers…but with online learning ...there's a lot of anxiety there because now they're not friends with the people they are learning with." One apprehension that teachers had about the reduction of content was the impact this would have on licensure. One Psychology teacher shared that, "If you want to take the licensure exam for psychometricians, you need this course. [Since] we weren't able to cover all the topics... [students] have to catch up [themselves]." One Chemistry teacher echoed the sentiment, "...you cannot get your Chemistry license if you haven't physically had this class to the degree that they dictate. This means that when we go back to school, they have to go back and do the labs because the law says they have to."

While many students and teachers welcomed the decision to give all students a passing grade, many teachers continued to offer learning opportunities at no risk. However, students generally did not take advantage of these. One teacher observed that, prior to the April 7 memo, many students continued to participate in her class discussions. After the memo was issued, very few students persisted because they saw classwork as optional. Major requirements had to be scrapped including final oral presentations, poster presentations, and term papers. Thus, many teachers felt their students received a lower-quality learning experience during ERT compared to the first part of the semester.

4. Conclusion

Beyond the shift from face-to-face to online learning, ERT also marked a shift in educational approach. A pedagogy of care became the order of the day. Within the ADMU in particular, the faculty and administration put a greater emphasis on *cura personalis*, or *respect for all that makes up each individual* (Otto, 2013). This necessary choice was enacted through drastic reductions in content, the use of diverse delivery methods, the use of a range of communication platforms, and greater leniency in assessment. The choice also came with a cost: that of academic rigor. Under the circumstances, however, the tradeoff was necessary for the public good. When an institution notices uncertainty and anxiety in its members, it needs to feel with its members and respond accordingly. Such moves spring from elements of that institution's culture, values, and beliefs. In the case of the Ateneo, this was elevated to the institutional level and helped its members realize that *cura personalis* was not only care for an individual, but care for everyone.

Acknowledgements

We thank the teachers and students for their participation; Ma. Rosario Madjos for her assistance; Dr. Silvia Bartolic and Dr. Neil Guppy of the University of British Columbia for initiating the multinational, multi-institutional study of the COVID-19 Transition to Online Instruction; other universities in the consortium; and the Ateneo de Manila's University Research Council for the Off-Cycle COVID-19 Research Grant titled "*Lessons for Higher Education from the COVID-19 Transition to Online Teaching and Learning*".

References

Bartolic, S., Boud, D., Agapito, J., Verpoorten, D., Williams, S.,Lutze-Mann, L., Matzat, U., Moreno, M., Polly, P., Tai, J., Marsh, H.L., Lin, L., Burgess, J., Habtu, S., Rodrigo, M., Roth, M., Heap, T. and Guppy, N. (2021). A multi-institutional assessment of changes in higher education teaching and learning in the face of COVID-19. *Educational Review*.

Baker, V. L. (2020, March 25). How colleges can better help faculty during the pandemic. *Inside Higher Ed*. https://www.insidehighered.com/views/2020/03/25/recommendations-howcolleges-can-better-support-their-faculty-during-covid-19

Bates, T. (2020, April 7). What should we be doing about online learning when social distancing ends? *Online Learning and Distance Education Resources*. https://www.tonybates.ca/2020/04/07/what-should-we-be-doing-about-online-learningwhen-social-distancing-ends/

Bozkurt, A., Jung, I., Xiao, J., Vladimirschi, V., Schuwer, R., Egorov, G., ... and Paskevicius, M. (2020). A global outlook to the interruption of education due to COVID-19 pandemic: Navigating in a time of uncertainty and crisis. *Asian Journal of Distance Education, 15*(1), 1-126.

Coutts, C. E., Buheji, M., Ahmed, D., Abdulkareem, T., Bujehi, B., Eidan, S. and Perepelkin, N. (2020). Emergency remote education in Bahrain, Iraq and Russia during the COVID-19 pandemic: a comparative case study. *Human Systems Management 39*, 473 – 493. https://DOI 10.3233/HSM-201097

Frost, P.J., Dutton, J.E., Mailis, S., Lilius, J.M., Kanov J.M. and Worline, M.(2006). Seeing Organizations Differently: Three Lenses on Compassion. *The SAGE Handbook of Organization Studies*, 2nd Ed. CA: SAGE Publications Ltd. DOI:http://dx.doi.org/10.4135/9781848608030

Gelles, L. A., Lord, S. M., Hoople, G. D., Chen, D. A., and Mejia, J. A. (2020). Compassionate flexibility and self-discipline: Student adaptation to emergency remote teaching in an integrated engineering energy course during COVID-19. *Education Sciences, 10*(11), 304.

Hewson, C. (2015). Research methods on the Internet. *Communication and Technology*, 5th Ed. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110271355-016

Johnson, N., Veletsianos, G., and Seaman, J. (2020). U.S. faculty and administrators' experiences and approaches in the early weeks of the COVID-19 pandemic. *Online Learning, 24*(2), 6-21. https://doi.org/10.24059/olj.v24i2.2285

Kanov, J.M., Laitlis, S., Worline, M.C., Dutton, J.E., Frost, P.J. and Lilius, J.M. (2004). Compassion in organizational life. *American Behavioral Scientist, 47*(6), 808-827. https://doi.org/10.1177/0002764203260211

Otto, A. (2013). Cura personalis. *IgnatianSpirituality.com*. https://www.ignatianspirituality.com/cura-personalis/

UNESCO. (2020 April 20). *Dealing with Obstacles to Distance Learning*. https://en.unesco.org/news/dealing-obstacles-distance-learning

Veletsianos, G. and Kimmons, R. (2020, April 6). What (some) students are saying about the switch to remote teaching and learning. *Educause Review*. https://er.educause.edu/blogs/2020/4/what-some-students-are-saying-about-the-switch-toremote-teaching-and-learning.

Facilitating Collaborative Learning among Businesses, Faculty, and Students in a Purely Online Setting

Joseph Benjamin ILAGAN*, Matthew Laurence UY, Vince Nathan KHO & Joselito OLPOC
Department of Quantitative Methods and Information Technology, Ateneo de Manila University, Philippines
*jbilagan@ateneo.edu

Abstract: Collaborative learning is a situation where two or more people attempt to learn together. We explain how we designed and facilitated collaborative learning among businesses, faculty, and students in a purely online setting during strict lockdowns during COVID-19. The design follows the four areas involving successful collaborative learning: initial conditions, collaboration roles, the scaffolding of interactions, and interaction monitoring and regulation. The course followed a blend of professional consulting engagements, student internships, and faculty externships. The primary discipline serving as the basis for the consulting engagements is business analytics, which covers various computer programming, statistics, and data visualization skills. While the overall program spans multiple academic terms, this paper focuses on the pilot term consisting of chosen business management students interested in programming and analytics. Despite logistical challenges and apprehensions among student participants in the middle of the term, the results were in line with generally expected learning outcomes based on feedback from the participants.

Keywords: Collaborative learning, asynchronous learning, online learning, COVID-19, blended learning, industry-university best practices, cognitive apprenticeship

1. Context and Motivation

Collaborative learning is when two or more people attempt to learn together (Dillenbourg, 1999). A related concept is a social aspect of learning (Vygotsky, 1997) in that learning is a consequence of social interaction. The Ateneo de Manila University in the Philippines has been pursuing opportunities for industry-university collaboration as it sees several advantages. Such partnerships are even more imperative in business analytics, where high-level competencies aligned with interdisciplinary, real-world market requirements are essential (Wang, 2015).

Like internships, the students will gain applied competencies in the real world, but with the presence of teachers and business clients working with them (Neumann & Banghart, 2001). The setup will allow the transfer of soft skills from businesses and teachers with industry experience to students and more junior faculty. Business faculty will have continuous involvement with industry, which will help contextualize lessons they teach in major business subjects. Thesis and Capstone mentoring and guidance will be richer and more up-to-date with the constant exposure to industry engagements. Neumann and Banghart (2001) refer to this setup as an externship. Businesses have access to academic (student and teacher) expertise otherwise not found in the regular job and professional consulting market. Ultimately, however, businesses will welcome collaboration with academia to achieve business impact—how the newfound knowledge from the collaboration affects company performance rather than activity output (Greitzer et al., 2010). This paper will interchange the terms business, company, client, business partner, data partner, and data client.

The industry-university collaborative learning setup takes advantage of the benefits of social learning (Vygotsky, 1997), this time across three groups: students, faculty, and industry. Neumann and Banghart (2001) describe the concept of consulternships, which reflect a blend of professional consulting engagements, student internships, and faculty externships. Another term for this type of collaboration setup with corporations of various sizes is Knowledge Transfer Partnerships (KTP)

(Roulston & McCrindle, 2018). In traditional internships, most interns or apprentices are often entry-level appointments involving menial tasks. Businesses prefer to place people already with experience in more critical positions. The task is often perceived as "boring" and can leave the students isolated and discouraged (Neumann & Banghart, 2001).

This paper tackles two research questions: 1) What class structure is needed for this type of collaborative setup? 2) What roles do faculty, students, and business partners need to play to make this setup successful? This paper describes the researchers' steps in designing and facilitating an online course on business analytics with collaborative learning (Dillenbourg, 1999) as the primary theoretical framework to increase the probabilities of interactions for positive learning outcomes. Complementary structural support in designing the remote and online course comes from the theory of Transactional Distance (Moore & Kearsley, 2011). Cognitive apprenticeship (Bransford, Brown, & Cocking, 2000, as cited in Ghefaili, 2003) also comes into play in shaping instruction design for this collaboration setup. Last, Self-Efficacy (Bandura, 1977), scaffolding, and learning involving social interaction (Vygotsky, 1997) help address gaps in what students could not learn if left on their own and what roles faculty need to play. Lastly, where applicable, concepts from cooperative learning (Johnson & Johnson, 2002; Gillies, 2016) will support learning independently, sharing of resources, and in small groups.

2. Program and Course Design

This paper focuses on a pilot class under a larger data partnership collaboration initiative by the Ateneo de Manila University (ADMU) in the Philippines. The course, Applied Business Analytics, involved two sections. One section had 28 students, while the other had 35 students. All students were from the Bachelor of Science in Management Engineering Program from the John Gokongwei School of Management of the ADMU, ranging from 18 to 21 years old. There were 33 male and 30 female students. All the students already took the required math and programming subjects before the class and some majors in strategy and decision science. The study works with a sample representing the whole batch of BS Management Engineering Students at the Ateneo de Manila University as the population. The faculty picked the students to attend this pilot class based on past interactions and observations of skills and attitudes. The criteria were based on past grades in programming classes and extra-curricular activities related to programming and analytics. Based on informal and anonymous surveys conducted, most of the students did not have previous analytics-related experiences or engage directly with business clients or employers. The class ran for eight (8) weeks, but the first week was spent on introductions and class administrative matters.

2.1 Challenges

The undertaking of an industry-university collaboration at this time presents many unknowns. All parties (students, teachers, and business clients) will have to fill specific learning gaps.

The University continues to conduct learning activities remotely and online due to the worsening COVID-19 situation. Traditional collaborative learning involves being at the same physical location (Dillenbourg, 1999), but the hope is that technologies available for online collaboration will make this physical distance less of an issue.

Transactional distance (Moore & Kearsley, 2011) refers to physical (especially in distance learning), pedagogical and psychological gaps, particularly between instructor and student and among students. The wider the gap, the more negative the impact on learning. In this class, transactional distance also includes gaps with industry contacts. Transactional distance involves three dimensions: structure, dialogue, and autonomy (Moore & Kearsley, 2011). Since not all students will have the same level of capacity for self-management, the overall design of the course needs the right amount of structure and dialogue. And, because it is impossible to provide a predefined structure in this setup, a lot of dialogue is an integral part of the course design. Teachers play a key role in guidance and constant assurance.

Even with the handpicked students, not all of them have a sufficient level of self-autonomy (Moore & Kearsley, 2011) and self-efficacy (Bandura, 1977). Overcoming transactional distance is nothing new in the university (Ilagan, 2020). However, this new program introduces unprecedented situations involving teachers, students, and business partners in a class setting.

Teachers and students had gaps in skills and experience in a natural business setting. Due to this uncertainty, the students had difficulty gauging whether or not they were on the right track regarding their work. Some teachers were more familiar with the hard math and technology required, while others would have the industry exposure. There is a need for more than one teacher in any meeting with a client due to the interdisciplinary and diverse nature of business problems tackled, thus leading to potential faculty exhaustion and burnout. Due to this uncertainty, the students had difficulty gauging whether or not they were on the right track regarding their work.

The business clients, too, don't have all the skills needed to help make this collaborative learning setup succeed. One serious gap is the understanding of their business requirements, though this varies across businesses. The clients generally know what data they have, but they don't know what they can do.

Unlike in a classroom setting, there is ambiguity in this new setup. With students, teachers, and business clients having skills and expectations gaps, there will inevitably be confusion and struggle in coming up with the expected insights. A mitigating strategy should follow an iterative process, starting with exploratory data analysis and prototyping (Wirth & Hipp, 2000), complemented with constant feedback from the clients.

2.2 Expected Learning Outcomes

Unlike regular classes, assessment and grading follow a rubric based on specific behaviors rather than students getting the correct answer. To ease students' fear of teachers' subjectivity in grading, the students will submit a self-assessment report, but the teachers still provide the final grade.

Given the skills gaps among all the participants, the lack of clarity in business goals for the engagements, and the free-form nature of the collaboration, there is the need for emotional support from all participants throughout the process. Because of many unknown aspects, students have to do self-study, research, data gathering, and inquiry independently (Roulston & McCrindle, 2018). Applications of theoretical material in real-life scenarios make content easier to understand, while the real-life application demonstrates the relevance of content (Roulston & McCrindle, 2018).

2.3 Class Design and Management

This subsection addresses the first research question of this paper: What class structure is needed for this type of collaborative setup? Dillenbourg (1999) offers four classifications of ways to increase the probability that some types of learning interactions occur: initial conditions, collaboration roles, the scaffolding of interactions, and interaction monitoring and regulation.

The skills gaps described earlier involve inert knowledge, and cognitive apprenticeship (Collins, Brown, and Newman, 1989, p. 453) is one way to address this. One of the goals of cognitive apprenticeship is to make the thinking processes of a learning activity visible to both the students and the teacher. The teacher can then employ the methods of traditional apprenticeship (modeling, coaching, scaffolding, and fading) to effectively guide student learning (Collins et al., 1991).

The design of this class touches on factors of successful Business Analytics programs (Wang, 2015). These factors include interdisciplinary collaboration with other departments or industries, aligning courses with the practice's needs, exposing students to real-world projects and industry professionals, blending statistics and quantitative methods, and strengthening the faculty's expertise.
The first three factors directly benefit from collaborative learning with industry participants. However, coordination with other departments did not happen for the pilot class due to lack of time.

2.3.1 Initial Conditions

Seven (7) business clients signed up for the data analytics partnership with the University. The recruitment and invitation of the business partners proceeded organically, with contacts of teachers from the alumni and business network becoming prime candidates. In searching for business partners, the program planners considered the challenges related to data availability and the feasibility of the expected scope of work, which became the basis for targeting appropriate businesses. Some partners had internal data privacy and information security policies that limited the data that they could provide for this project. Other times, partners or students did not have the means to collect the data needed for the project.

Teachers and students playing the roles of account manager and lead consultant would have to set expectations with business clients and avoid overpromising. One way of preventing overcommitment is to focus on impactful but not necessarily urgent or business-critical work.
After finalizing the data partners for the course, planning class content followed their business needs. Some business requirements needed not previously taken by the students in their earlier classes.

The designated project manager must update the business partners from time to time to make them feel involved and know about any team task problems. Regular updates done iteratively would avoid significant surprises in the end.

2.3.2 Set Collaboration Contracts Based on Roles

This subsection addresses the second research question of this paper: What roles do faculty, students, and business partners need to play to make this setup successful? The University team prepared a generic client engagement framework that outlined roles and responsibilities on both sides of the client and the University in more detail. For the primary roles, a person may assume one or more of these. Faculty act as account managers, project managers, and lead consultants. The students act as associate consultants, data engineers, and analysts. The business client serves as the subject matter expert and the recipient of any recommendations and actionable insights from faculty and students.

The business client's needs may force faculty and students to focus on consultant and project management roles. However, students still expect teachers to perform content delivery-related tasks.

2.3.3 Scaffolding Productive Interactions

In learning, a study (Smith and Ragan, 2004) defines scaffolding as cognitive processing support that the instruction provides learners. This concept originated from Vygotsky's sociocultural theory (Vygotsky, 1978, as cited by Schutt, 2003) and the notion of the zone of proximal development (ZPD). The ZPD is the gap between the learner's accomplishment versus what would have been possible with an expert, teacher, or a more competent person present.

Sample programming code comes in the form of work already created by teachers and the student core technical team. Other code snippets are carefully curated from multiple sources from the Internet. One class had weekly lectures on how to apply data science around different problems. For client interaction scaffolding, one faculty member acts as the lead consultant to facilitate meetings with the clients for client interactions. Students get to observe how to handle real-world client interactions. These observations will serve as models in the future (Bandura, 1977, as cited by Hodges, 2008).

2.3.4 Monitoring and Regulation of Interactions

With the online setting, monitoring and project management could only be possible through technologies already available to all parties. Internal online chat and audio meetings took place through Discord, and its use is nothing new (Kruglyk, Bukreiev, Chornyi, Kupchak, & Sender, 2020). For recorded video meetings and ad-hoc lectures, the teachers offered their Zoom accounts. For external meetings, the default preference of all parties (including the business partners) was Zoom. However, some organizations have standardized the use of Microsoft Teams.

Informal pulse checks with the classes and across project groups continued throughout the pilot period via Discord. In addition, the teachers prepared a more structured pulse check in the form of a survey mid-way through the term. A final survey was conducted a few weeks after the end of the class.

3. Results

Before the class, students were generally happy but fearful of not doing well because they lacked the business analytics and customer engagement experience. This was also the time when teams were still determining what outputs would make sense to the business clients. While the class covered general analytics frameworks, the content was also motivated by the project requests of the clients. The students enriched their knowledge through research and self-study on specialized topics.

During the class, the students felt overwhelmed because they realized how much they still need to learn, but they were satisfied with the flexibility of the structure and the guidance they received from the faculty. The students were also satisfied with the ad-hoc workshops, and heavy lifting is done by faculty and the student core technical team. For soft skills in handling general client communication, scheduling meetings, business presentation, and account management, faculty managed most client-facing activities at the start. Teachers worked on programming examples through Python Jupyter notebooks. A large part of the learning experience involved regular personalized consultations with project teams about approaching their specific projects.

After the class, the students were surprised by what they had been able to do. They were also happy that they were able to satisfy the business partners they served. Some said that this had been the best class in their stay in college but that the class should have been done in a regular (16-week) semester instead of the short eight-week quarter. And, because of the short time allocated, the students felt they didn't have enough opportunities to fill particular skills gaps. Despite everything, they acknowledged that it is normal to have these knowledge gaps. They understand that how to address such gaps is part of the intended learning outcomes. In general, students had also expressed feeling more confident in applying different analytics tools after the course compared to how they felt before the course. Teachers had to exert more effort to deliver additional classroom material, manage project groups, and communicate with the business clients, thus making them feel exhausted. However, the teachers also were satisfied with the learning outcomes exhibited in the real-life business impact of the projects. Finally, business clients expressed how much they learned from the engagement. They shared that the students' recommendations were presented to upper management as well.

4. Discussion

Since everything has been experimental, the collaboration program will benefit from documenting the journey. Documentation may initially be done informally through Discord conversations and then come up with a reflection towards the end of the class. One goal is to produce artifacts (templates, reusable code, video tutorials, and others) to benefit future batches taking this class. Teachers and business partners do not have to re-establish business objectives from scratch if prospective students can reference the outputs from before. This paper may serve as the basis for a more generic model involving longer-running thesis projects and more impactful internship and externship arrangements with business partners. However, the exhaustion experienced by faculty is a significant risk that needs to be taken into account as this initiative scales up to similar but larger collaborative projects. Finally, the University needs to find ways to quantify the success of the collaboration effort with industry partners. Success metrics will relate to learning outcomes for the University. It will align with business impact for industry partners (Greitzer, Pertuze, Calder & Lucas, 2010).

Acknowledgements

The course designed and delivered and the course that served as the basis for this paper were made possible by the Ateneo de Manila University, the Ateneo's Department of Quantitative Methods and Information Technology, and the John Gokongwei School of Management of the Ateneo. Special thanks to Ma. Mercedes T. Rodrigo for the guidance and mentorship throughout the whole process until the writing of this paper. We thank them all for their cooperation and support.

References

Andrews, J., and Higson, H. (2008). Graduate Employability, ‘Soft Skills’ Versus ‘Hard’ Business Knowledge: A European Study. *Higher Education in Europe*, *33*(4), 411–422. https://doi.org/10.1080/03797720802522627

Bandura, A. (1977). Self-efficacy: Toward a Unifying Theory of Behavioral Change. *Psychology Review*, *84*(2), 191–215.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., and Krathwohl, D. R. (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1: Cognitive Domain.

Bransford, J., Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school*. National Academies Press.

Collins, A. (1991). Cognitive apprenticeship and instructional technology. *Educational values and cognitive instruction: Implications for reform*, 1991, 121-138.

Dillenbourg, P. (1999). What do you mean by collaborative learning? *Collaborative-Learning: Cognitive and Computational Approaches*, 1–19.

Gillies, R. M. (2016). Cooperative learning: Review of research and practice. Australian Journal of Teacher Education (Online), 41(3), 39-54.

Greitzer, E. M., Pertuze, J. A., Calder, E. S., & Lucas, W. A. (2010). Best practices for industry-university collaboration. *MIT Sloan Management Review*, *51*(4), 83.

Hodges, C. B. (2008). Self‐ efficacy in the context of online learning environments: A review of the literature and directions for research. *Performance Improvement Quarterly*, *20*(3‐ 4), 7-25.

Ilagan, J. B. (2020). Overcoming transactional distance when conducting online classes on programming for business students: a COVID-19 experience.

Johnson, D., & Johnson, R. (2002). Learning together and alone: Overview and meta-analysis. Asia Pacific Journal of Education, 22, 95-105.

Ghefaili, A. (2003). Cognitive apprenticeship, technology, and the contextualization of learning environments. *Journal of Educational Computing*, Design & Online Learning, 4(1), 1–27.

Kruglyk, V., Bukreiev, D., Chornyi, P., Kupchak, E., and Sender, A. (2020). Discord platform as an online learning environment for emergencies. *Ukrainian Journal of Educational Studies and Information Technology*, *8*(2), 13–28. https://doi.org/10.32919/uesit.2020.02.02

Moore, M. G., & Kearsley, G. (2011). *Distance education: A systems view of online learning*. Cengage Learning.

Neumann, B. R., & Banghart, S. (2001). Industry‐ university “consulternships”: an implementation guide. *International Journal of Educational Management*.

Roulston, D., and McCrindle, R. (2018). Engaging students in research with ‘real-world’ outputs: Making an impact outside of the lecture theatre. *Shaping Higher Education with Students*, 208–221.

Schutt, M. (2003). Scaffolding for Online Learning Environments: Instructional Design Strategies that Provide Online Learner Support. *Educational Technology: The Magazine for Managers of Change in Education*, *43*(6), 28–35.

Smith, B. L., & MacGregor, J. T. (1992). What is collaborative learning. *Towards the Virtual University: International Online Learning Perspectives*, 217-232.

Smith, P. L., & Ragan, T. J. (2004). *Instructional design*. John Wiley & Sons.

Vygotsky, L. S. (1978). Socio-cultural theory. *Mind in society*, *6*, 52-58.

Vygotsky, L. S. (1997). *The collected works of LS Vygotsky: Problems of the theory and history of psychology* (Vol. 3). Springer Science & Business Media.

Wang, Y. (2015). Business intelligence and analytics education: Hermeneutic literature review and future directions in is education. *Proceeding of Twenty-First Americas Conference on Information Systems (AMCIS), Puerto Rico*.

Wirth, R., & Hipp, J. (2000). CRISP-DM Towards a Standard Process Model for Data Mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 29–39. London, UK: Springer-Verlag.

Educational Leadership and Children's Resilience: German and Polish Schools during COVID-19

Paulina BURKOT*, Amy SEPIOŁ & Nataliia DEMESHKANT
Pedagogical University of Cracow, Poland
*paulina.burkot@student.up.krakow.pl

Abstract: As a result of the pandemic, education has moved into virtual space on an unprecedented scale. The purpose of this study was to determine the influence of the COVID-19 pandemic on education and to compare the problems and ways of coping with them between Polish and German schools. The study used a qualitative measure, which was realized through telephone interviews with school principals. The received responses have been transcribed and categorized. The final stage was a comparative analysis of the data obtained. The results of the study revealed that the most common problems concerned: the possession and functionality of teaching resources for remote learning (computer equipment, network connection), ensuring the needs of face-to-face relationships (lack of peer meetings and no direct contact with the teacher), mental problems caused by changes in education (depression, anxiety, feelings of stress) in teachers and students. Among the ways schools deal with these problems are leadership role of school leaders (support, staff training, mental support of students and teachers (meetings with pedagogue, school psychologist, conducting remote social meetings with students).

Keywords: Online learning, COVID-19, educational problems

1. Introduction

The pandemic significantly influenced many aspects of people's life, ranging from economic processes to education. In March 2020 governments started implementing certain measures in order to stop spreading the COVID-19 disease by closing schools and switching into online teaching. Education in many countries faced a very difficult challenge to educate the youth from the distance using modern computer technology. The closure of educational establishments has suddenly disrupted the school life of many students and their families which caused long-term consequences regarding social ties and mental health (Rajabi, 2020).

Teachers, who have developed their teaching and training skills based on direct contact with their students inside the classroom throughout the years, had to resign from it and change it into digital technology. Until that time, they had been using it to enrich and diversify the lesson and not as a constant part of the educational process. Studies show that only 15% of Polish teachers had had an experience in distance teaching and 85% had not had any experience in distance teaching before the COVID-19 epidemic started (Buchner et al., 2020). Education has been transformed into an online mode, on an untested and uncommon scale (Burgess et al., 2020).

After the first stage of joy, connected with the announcement of schools' closure, children and teenagers experience many emotions and problems typical for a crisis situation (Pyżalski, 2020). Some students suffer from depression, caused by isolation, loneliness and lack of contact with friends. Younger children have trouble accepting the new situation, they are not able to understand it completely, they are confused and lost. The response to such situation is nonroutine and dependent on many factors, for example affective responsiveness, past experiences, opportunity and ability to use the social support (of family and friends). Children's and teenagers' mechanism of working through the crisis is less developed. Therefore, they belong to a group which is the most vulnerable (Pyżalski, 2020). A vital concern about students' wellbeing was presented in a number of recent studies (Hamilton et al., 2020; Ferri et al., 2020; MacDonald & Hill, 2021). Recent studies highlighted that childhood education, due to its connection to everyday and real-life experience, as well as children's greater dependence on parents, faces particular challenges when moved online (Spiteri, 2021) and expresses concern about children's socialization (MacDonald and Hill, 2021).

A huge responsibility as well as pressure fell on principals' shoulders. They had to ensure the safety and psychological support of pupils, teachers and change the functioning of educational institutions. Managing crisis and change are now essential skills for a school leader. Running an effective school in challenging times requires more than routine problem solving (Harris, 2020).

School leaders engage in a diverse set of tasks related with teaching, learning, and assessment for promoting students' learning experiences and achievements, promoting and supporting continuous staff development, planning and monitoring available resources (including infrastructure), meeting the external accountability requirements, and cultivating a nurturing school culture with the extended school community (Swan Dagen et al., 2017).

School leadership is considered a core aspect of a successful school improvement across the world, especially in the light of the emerging paradigms of increased school autonomy and accountability (Sergis et al., 2018).

2. Materials and Methods

The survey was conducted at the end of the year 2020 on a group of 26 respondents (German =13, Polish =13), which were principals and managers of primary and secondary schools in Poland and Germany. Participants for interviews were chosen by random selection and their own willingness for the interviews. They were informed that the interview would be recorded and analyzed for scientific purposes. Qualitative measure was used in the study conducted by phone interview with school leaders. The duration of each interview oscillated between 20-30 minutes. The participants' responses were recorded. After all the interviews, the interviews were transcribed and translated into English. A list of codes was then created to match the respondents' responses. When analyzing the results, the most frequently repeated responses were assigned the appropriate code key categories. Each participant answered 12 open questions concerning pandemic risks, methods of preventing and coping with them, as well as the principal's system of work in the difficult pandemic period. These questions focused on three main research questions:

1. What are the risks during the COVID-19 pandemic for students from the perspective of school leaders? Including own resilience.
2. How can school leaders support children's resilience during the COVID-19 pandemic? Learning and teaching issues.
3. What (non-/official) strategies do German and Polish school leaders use in their schools?

According to the research questions above, three leading categories were proposed: Changes in the way of teaching, Student's resilience, and Principal's leadership. The list of subcategories regarding each category was conducted (Table 1 shows categories, subcategories, and examples of answers).

Table 1. *Categories and Codes from Interviews*

Categories	Subcategories	Examples of answers
Changes in the way of teaching	Complexity of new regulations	I like online staff meetings very much. Online meetings are very comfortable. Use technology more often.
	Most challenging part of remote education	Trouble with conducting lessons with students in the classroom and online. Fear of showing one's face to camera. Lack of routine (going to school) generates many behaviors in students.

		Problems with isolation at home. Teachers were not prepared for remote teaching.
	Most frequent problems in remote education	Technological issues. Little visibility of students. Stability of internet connection. Access to laptops. Teachers focusing on other things instead of teaching.
Student's resilience	Coping with stress and problems	Organizing various meetings for staff and students. Talking with each other. Meetings with specialists.
	Supporting students	Helping children integrate. Meetings with specialists. Asking students about their well-being.
Principal's leadership	Decreasing the negative impact of pandemic	Be in touch with parents and students. Meetings with teachers individually. Suggest visiting school and talking with someone or a specialist. Important on focusing on relationships with people.
	Concepts after pandemic	Online meetings are very comfortable. Use technology more often.
	Guidance or regulation from government	Local government did very well- we got what we asked for. It would be perfect to have a calendar when something starts and finishes. We really appreciate funding and teacher training.

The final stage of the study was a comparison of the data obtained and categorized between Polish and German schools, which allowed to determine similarities and indicate differences.

3. Results

For the purposes of this article, our main focus is on three key common themes which were generated from the data with regard to school leadership and children's resilience during remote education. Obtained data was analyzed according to the three main categories in the comparison between Polish and German schools.

Table 2 presents a list of similarities between the two countries in terms of educational challenges, student's resilience and principal's leadership.

Table 2. *List of Similarities between Schools in Poland and Germany*

Categories	Similarities in both countries
Changes in the way of teaching	Technological problems (access to computers, laptops, tablets, the Internet). Lack of adequate preparation of the teaching staff for remote classes. Organization of lessons from the teachers' perspective more difficult and time-consuming.

	Digital fatigue both among students and teachers. Increased contact between teachers and students after lessons.

	A complete change in teaching methods.
Student’s resilience	Psychological and emotional problems among many students (depression, anxiety). Initiatives, classes, teacher talks aimed at supporting students.
Principal's leadership	Recognizing the advantages of digital technologies and the willingness to introduce some technological solutions after the pandemic ends. Psychological support for students and school staff by principals. Mutual cooperation between teachers and principals.

The similarities within the 3 major categories were very similar. Many of the educational problems were repeated, regardless of the country. Depending on the school and the country, different methods were developed to solve the problems that occurred. Table 3 lists the opinions of the participants in this study regarding the differences between the countries during the COVID-19 pandemic.

Table 3. *Summary of Differences between Schools in Poland and German*

Differences in both countries		
Categories	Poland	Germany
Changes in the way of teaching	Average effectiveness of remote education	High effectiveness of remote Education
	Striving to the realization of the core curriculum	Adjusting the teaching content to the realities of the pandemic, flexibility of actions
Student’s resilience	Focus on mutual understanding and help	Focus on various sources of stress and problems
Principal's leadership	Support from specialists, pedagogues, psychologists	Relieving the stress level through various types of breathing techniques, mutual conversations
	Expressing a desire for a greater level of autonomy	Balance between the activities of the school principal and ordinances

The analysis of all subcategories related to the category “Changes in the way of teaching” allowed us to assume that teachers in both countries were not prepared to conduct remote lessons. Sudden decisions made by the government caused the increase in frustration because teachers were not able to properly prepare themselves to classes in such a short time. Such lessons required much more effort. Some schools have organized special trainings of how to use different tools in online teaching. However, many teachers had to train on their own how to use modern technology in education, in order to live up to parents’ expectations and not to lower the quality of teaching. In Polish schools, there was a strong emphasis put on the realization of the core curriculum to ensure that students were prepared for exams as well as possible. However, in German schools, lessons have been conducted in a more flexible way. The topics were organized in such a way that they could be covered online.

Regarding category “Student’s resilience”, the educational institutions in both countries have been taking different actions in order to improve students’ mental condition. In Polish schools, more emphasis was placed on supporting specialists, for example pedagogues and psychologists. Teachers organized teams to communicate with students. During homeroom hour they tried to talk with students about their well-being and in case of some worrying signals, be ready to provide assistance. On the other hand, in German schools, meditation techniques and breathing exercises have been used to reduce stress levels in students.

Data analysis in the category “Principal's leadership” showed that principals acted on the basis of government regulations, which in the case of Germany were short-term and rapidly transformed, along with the rapidly changing epidemiological situation, which was a major challenge, as well as the source of many frustrations. Spontaneous decisions made by the government decision-makers forced principals to act immediately with the lack of time to prepare reliably for their implementation. In Poland, principals acted in a similar way in implementing the regulations from the Ministry of National Education. Western neighbors had more support in making decisions from teachers, as well as more rules to make work easier. Polish school leaders received support from students’ parents and various NGOs. In addition, principals showed a great desire to have greater autonomy. They wanted the government to allow them to decide if they should close the school, because each institution is different and has its own specifics. German principals, on the other hand, would prefer a better balance in terms of government regulations and their own decisions. In both countries, most schools had mutual cooperation and communication between educational institutions.

4. Discussion

Teachers not only in Poland and Germany, but also in most countries, were not prepared for such a sudden change in education, both in terms of mental, technological and methodical preparation. Teachers have taken the trouble to transfer teaching content and materials to virtual space and to acquire competences in using the necessary software (Allen et al., 2020). The pandemic has shown what the consequences will be if schools do not keep up with the fundamental process of ICT transformation. It will therefore be crucial to provide teachers with opportunities for professional development and training for future teachers (Daniel, 2020). The most common factors limiting remote education were concerns about ensuring fair teaching for all students, hardware shortages among students, and the lack of internet access (Hamilton, 2020).

Our research on educational leadership and psychological resilience of students in schools in Poland and Germany has shown that the pandemic has clearly had an impact on the mental health of pupils in both countries. Previous studies confirmed that outbreaks of infectious diseases have a negative impact on the health and mental condition of the society (Sim et al., 2010). Children and adolescents are more susceptible to the psychological effects of the COVID-19 pandemic, they tend to show more negative psychological effects, so they need psychological support from three cooperating systems: the social system, the school system and the family system (Zhou, 2020). Recent study stressed on significant psychological problems of older children and teenagers as a result of imposed social isolation. There has been a development of anxiety, fear, somatic symptoms, sleep disorders, depression, feelings of anger and irritability, regret and loss, as well as post-traumatic stress (Esposito et al., 2021). Our results were almost identical. Students often pointed out how much they longed to meet their friends and teachers (Ewing et al., 2021). Our results showed that most schools have taken initiatives to support pupils during isolation by organizing special meetings for students, usually during the homeroom hours, but not only. Almost all teachers kept in touch with students and their parents (König et al., 2020).

Trust has been identified as a key measure of crisis leadership effectiveness. It requires authenticity and honesty on the part of the leader (Schoenberg, 2005). In addition, the pace of change in this pandemic is unprecedented, so school leaders will need to be involved in ongoing crisis and change management, which will require support and collaboration from all staff (Harris, 2020).

5. Conclusion

The results of the study made it clear that, regardless of the country during the COVID-19 pandemic, there were many similarities in the problems related to remote education, teacher preparation for online classes, mutual cooperation, and the deterioration of the well-being of pupils in both countries due to isolation. At the same time, some differences have been revealed between schools in Poland and Germany concerning, among others, the evaluation of remote education, the realization of the content of the core curriculum, as well as the sources of support for school principals at this difficult time.

References

Allen J., Rowan L., Singh P. (2020). Teaching and teacher education in the time of COVID-19. *Asia-Pacific Journal of Teacher Education* Volume 48, 2020. https://doi.org/10.1080/1359866X.2020.1752051

Buchner A., Wierzbicka M. (2020). Edukacja zdalna w czasie pandemii. Edycja II. Centrum Cyfrowe, https://centrumcyfrowe.pl/wp-content/uploads/sites/16/2020/11/Raport_Edukacja-zdalna-w-czasie-pandemii.-Edycja-II.pdf.

Burgess, S. and Sievertsen, H. (2020). Schools, Skills, and Learning: The Impact of COVID-19 on Education.

conversation with Virginia Braun and Victoria Clarke about thematic analysis. Qualitative Research in

Daniel S. J. (2020). Education and the COVID-19 pandemic. Prospects, 1–6. *Advance online publication*. https://doi.org/10.1007/s11125-020-09464-3

Esposito, S., Giannitto, N., Squarcia, A., Neglia, C., Argentiero, A., Minichetti, P., Cotugno, N., & Principi, N. (2021). Development of Psychological Problems Among Adolescents During School Closures Because of the COVID-19 Lockdown Phase in Italy: *A Cross-Sectional Survey. Frontiers in pediatrics*, 8, 628072. https://doi.org/10.3389/fped.2020.628072

Ewing, L. A., Cooper, H. B. (2021). Technology-enabled remote learning during COVID-19: perspectives of Australian teachers, students and parents, Technology, Pedagogy and Education, DOI: 10.1080/1475939X.2020.1868562

Ferri, F., Grifoni, P., & Guzzo, T. (2020). Online Learning and Emergency Remote Teaching: Opportunities and Challenges in Emergency. Situations. Societies 2020, 10, 86.

Hamilton, L.S.; Kaufman, J.H.; Diliberti, M. Teaching and Leading through a Pandemic: Key Findings from the American Educator Panels. Spring 2020 COVID-19 Surveys; RAND Corporation: Santa Monica, CA, USA, 2020.

Harris A., Jones M. (2020) COVID 19 – school leadership in disruptive times, School Leadership & Management, 40:4, 243-247, DOI: 10.1080/13632434.2020.1811479.

König, J., Jäger-Biela, D.J., & Glutsch, N. (2020). Adapting to online teaching during COVID-19 school closure: teacher education and teacher competence effects among early career teachers in Germany. *European Journal of Teacher Education*, 43, 608 - 622.

MacDonald, M.; Hill, C. The Educational Impact of the Covid-19 Rapid Response on Teachers, Students, and Families: Insights from British Columbia, Canada. Prospects 2021.

Psychology 1–22. DOI: 10.1080/14780887.2019.1670765.

Pyżalski, J. (red.) (2020). Edukacja w czasach pandemii wirusa COVID-19. Z dystansem o tym, co robimy obecnie jako nauczyciele. *Warszawa: EduAkcja*.

Rajabi, M. (2020). Mental health problems amongst school-age children and adolescents during the COVID-19 pandemic in the UK, Ireland and Iran: A call to action and research. *Health Promotion Perspectives*, 10(4), 293-294.

Schoenberg, A. (2005). Do crisis plans matter? A new perspective on leading during a crisis. Public Relations Quarterly, 50(1), 2–6.

Sergis, S., Voziki, T., & Sampson, D. (2018). School leadership: An analysis of competence frameworks. In Digital Technologies: Sustainable Innovations for Improving Teaching and Learning (pp. 3-25). Springer, Cham.

Sim, K., Huak Chan, Y., Chong, P. N., Chua, H. C., & Wen Soon, S. (2010). Psychosocial and coping responses within the community health care setting towards a national outbreak of an infectious disease. *Journal of psychosomatic research*, 68(2), 195–202. https://doi.org/10.1016/j.jpsychores.2009.04.004

Spiteri, J. (2021). Quality Early Childhood Education for All and the Covid-19 Crisis: A Viewpoint. Prospects. Swan Dagen A., Morewood A., & Smith M. L. (2017). Teacher leader model standards and the functions assumed by National Board Certified teachers. In The Educational Forum (Vol. 81, No. 3, pp. 322-338). Routledge.

Zhou X. (2020). Managing psychological distress in children and adolescents following the COVID-19 epidemic: A cooperative approach. Psychological trauma: theory, research, practice and policy, 12(S1), S76–S78. https://doi.org/10.1037/tra0000754

The Use of Video Conferencing Applications Facilitating Behavioral Engagement during Synchronous Learning in the Time of Pandemic

Mark Anthony R. ARIBON III
Ateneo de Manila University, Philippines
*mark.aribon@obf.ateneo.edu

Abstract: Behavioral engagement in an online setting can be described as an on-task performance done by the students (Moser & Smith, 2015). This refers to how active the students are during synchronous learning using the features of the video conferencing applications such as opening their camera, speaking through the microphone, screen sharing contents, and using the chat box to answer and respond to questions (Al Mamun et al, 2016; Park & Bonk, 2007). This study describes the use of video conferencing applications in facilitating behavioral engagement during synchronous learning. 462 Grades 11 and Grade 12 students and five (5) teachers were invited to answer a self-report questionnaire. The results indicated that teachers provided opportunities for behavioral engagement to happen during synchronous learning by making the features always available for the students to use. Further, the study showed a very weak relationship between the challenges encountered by the students and the students' perception of the use of video conferencing applications contradicting other literature suggesting that the two variables affect each other. Moreover, this study found out a moderately strong relationship between the frequency of use of the features and the students' perception on the use of video conferencing application during synchronous learning.

Keywords: Technology integration, behavioral engagement, Philippines, video conferencing application

1. Introduction

The sudden surge of Coronavirus Pandemic (COVID-19) in the Philippines has greatly affected the education system. In the past year, the Department of Education (DepEd) has been implementing various modes of learning to augment traditional face-to-face teaching and learning to resolve the needs and the challenges of the current situation in the education sector (DepEd, 2020), applying two modes of teaching and learning: Asynchronous and Synchronous.

Finol (2020) described asynchronous learning where teachers provide students the materials to use during learning, such as printed modules, educational videos, PowerPoint files, documents, etc. In short, students learn the same content at their own pace. On the other hand, synchronous learning is when teachers and students interact over a video conferencing application and participate in learning discussion mimicking a traditional face-to-face classroom (Masa, Diaz, Delgado, & Esteban, 2014).

Schools in the Philippines have utilized video conferencing applications, such as Zoom, Google Meet, and Microsoft Teams to conduct synchronous learning classes. However, internet stability, lack of digital resources of teachers and students, and lack of training and material production for teachers are major concerns when conducting synchronous learning (May, 2020; Bott, 2020). Additionally, students' disengagement to the learning discussion being a significant challenge for teachers and school administrators (Hartwell, 2017; Bernard et al, 2004).

Behavioral engagement in an online setting does not differ from behavioral engagement constructs in the classroom as articulated by Fredrick, Blumenfeld, & Paris (2016). Literature have defined behavioral engagement as an on-task behavior and performance which includes students utilizing the chat function, activating their microphone, opening their video camera, and using virtual raised hands to be recognized, making this type of engagement the easiest one to observe (Moser & Smith, 2015; Al Mamun et al, 2016).

Teachers may be able to integrate audio, video, chat rooms, and other student response systems which will make the students active participants in the learning discussion. However, some instances where teachers and students encounter technical problems, such as distorted audio and video quality, internet instability, and power outage.

In this study, the researcher described the use of video conferencing application facilitating students' behavioral engagement during synchronous learning of senior high school students at Polytechnic University of the Philippines (PUP), Manila during the 1st semester, AY. 2020-2021.

PUP is a state university in the Philippines that houses almost 70,000 students from its twenty-two campuses nationwide. It continues to become the Philippines' partner in improving the plight of the poor, developing students holistically, and alleviating poverty amongst its students and the marginalized community.

Specifically, this study sought to answer the following research questions:

1. Is there a significant relationship between teachers providing opportunities for behavioral engagement and the students' frequency of use of the features of video conferencing applications?
2. Is there a significant relationship between the challenges encountered by the students and their perception on the use of video conferencing applications during synchronous learning?
3. Is there a significant relationship between the frequency of use of the features and the perception of the students on the use of video conferencing applications?

2. Methodology

In this section, the target participants, instruments, and procedures conducted in the study will be discussed thoroughly.

Participants – this study consists of 462 (179 Grade 11, 283 Grade 12) Science and Technology, Engineering, and Mathematics (STEM) senior high school students (SHS) who took up General Chemistry 1 and General Physics 1 and the five (5) teachers handling these subjects.

Instrument – The researcher used the instruments provided by Abaci & Goodrum (2015). It was modified to fit the criteria for other video conferencing applications and to make it more specific for teachers and students. Two (2) versions of the questionnaire were produced – Teacher's Questionnaire and Student's Questionnaire. In the teacher's questionnaire, the availability of the features of the video conferencing applications and how frequent do students use these features during synchronous learning were asked. In the student's questionnaire, items such as the video conferencing application they use in class, the availability of the features of the video conferencing applications, frequency of use of these features, challenges encountered, and their perception on using video conferencing applications were also included.

Procedures – The researcher translated the questionnaire into a Google Form format since classes in PUP are conducted online. The researcher asked assistance from the PUPSHS administrators to disseminate the Google Form link to the teachers and to the students through their group chats, Facebook groups, or virtual classrooms. Informed consents were also obtained from the participants.

3. Results and Analysis of Data

3.1 Participant's Demographics

The researcher invited 462 SHS STEM students and their teachers to answer the survey questionnaire.

Table 1. *Summary of Participants' Demographics*

		Students		Teachers	
Video Conferencing Applications	Google Meet	242 (52.4%)		2 (40%)	
	Microsoft Teams	173 (37.4%)		2 (40%)	
	Zoom	47 (10.2%)		1 (20%)	
Device/s Used	Smartphone	289 (61.9%)	Often	0	Sometimes

and Frequency of Use	Tablet/iPad	45 (9.6%)	Never	1 (20%)	Rarely
	Laptop Computer	287 (61.7%)	Sometimes	5 (100%)	Always
	Desktop Computer	83 (17.8%)	Rarely	0	Rarely
Internet Connection	Yes	428 (92.6%)		5 (100%)	
	No	34 (7.4%)		0	

The participants mentioned that most of their synchronous learning classes are conducted using Google Meet (52.4%/40%) or Microsoft Teams (37.4%40%) video conferencing applications. Additionally, students often use their smartphones (61.9%) to attend the synchronous learning since this device is not only for educational purposes, but practically for personal use as well. Teachers use laptop computers (100%) to easily navigate the video conferencing application.

When asked about the available internet connection at home, 428 students (92.6%) and all teachers (100%) are subscribed to an internet service provider. The other students rely on their network's mobile data and from their neighbor's/relative's internet connection.

3.2 Teachers Providing Opportunities for Behavioral Engagement during Synchronous Learning and Students' Frequency of Use of the Features of Video Conferencing Applications

This section discusses the teachers providing opportunities for behavioral engagement to happen during synchronous learning. Thus, by making the features available to the students to use, they also allow opportunities for behavioral engagement to happen during synchronous learning.

Table 2. *Summary of the Availability of the Features of Video Conferencing Application as Provided by the Teachers during Synchronous Learning and the Students' Frequency of Use of these Features*

Features of Video Conferencing	Teachers					Students
	M	*SD*	Description	*M*	*SD*	Description
Group messaging	3.80	1.79	Often	3.68	1.19	Often
Screen sharing	4.60	0.89	Always	3.22	1.42	Sometimes
Recording of video and shared content	4.40	0.89	Always	3.29	1.37	Sometimes
Emojis/Virtual raised hands as signal	4.20	0.84	Always	3.38	1.23	Sometimes
Microphone	4.60	0.89	Always	3.97	1.37	Often
Video Camera	4.60	0.89	Always	3.38	1.23	Sometimes

Legend: 4.20-5.00 = Always, 3.40-4.19 = Often, 2.60-3.39 = Sometimes, 1.80-2.59 Rarely, 1.00-1.79 = Never

Teachers confirmed that the features of the video conferencing applications, such as screen sharing, microphone, and video camera (M=4.60; SD=0.89) are always available for the students to use during synchronous learning. Hence, teachers do allow their students to share their screens to present PowerPoint slides, especially when proposing a plan, presenting an idea, or the activities.

With 0.05 level of significance, the obtained p value=0.00, presenting a substantial relationship, therefore rejecting the null hypothesis that there is no significant relationship between the availability of the features of the video conferencing applications and the students' frequency of use of these features.

This result suggests that students have more opportunities to be active participants during synchronous learning by using communication channels, such as the microphone, group messaging, and screen sharing features. Additionally, researchers found that students make use of these features to interact and participate in the learning discussion by asking questions, sharing ideas, and discussing with their classmates (Andrew, Maslin, & Ewens, 2015; Hudson, Knight, & Collins, 2012). More so, these students tend to be motivated to learn and possess higher order thinking particularly when teachers embed interactive software, audio, and video in the learning discussion (Armstrong & Thornton, 2012).

Therefore, when teachers allowed their students to maximize the use of the features of the video conferencing applications, most possibly, students would be able to freely provide their thoughts and ideas on the topics being discussed and share their experiences.

3.3 *Challenges Encountered by the Students and the Students' Perception on the Use of Video Conferencing Applications during Synchronous Learning*

The stability of the internet connection is essential for the video conferencing application features to work properly, if not, certain technical difficulties may be encountered during the conduct of synchronous learning which may have an impact on students' learning and engagement.

Most students complained that they often experience lagging screen sharing (*M*=3.42; *SD*=1.10) during synchronous learning when teachers are moving to succeeding slides of their PowerPoint or when playing video. Other students sometimes encounter distorted audio (*M*=3.00; *SD*=0.98) and video (*M*=3.16; *SD*=1.07) and trouble joining the meeting (*M*=2.89; *SD*=1.08) possibly because of internet instability. More so, students rarely experience insufficient bandwidth (*M*=2.97; SD=1.21) and power outage (*M*=2.23; *SD*=0.99).

Table 3. *Summary of the Students' Perception on the Use of Video Conferencing Applications Facilitating Behavioral Engagement during Synchronous Learning*

Students' Assessment on the Use of Video Conferencing Applications	*M*	*SD*	Description
It helped me…			
to communicate with my teacher.	4.09	0.93	Agree
to collaborate with my classmates.	3.92	1.03	Agree
to feel a sense of community and social presence	3.79	1.07	Agree
to attend class meetings remotely.	4.16	0.94	Agree
to learn the course materials/content.	3.99	0.99	Agree
to study for quizzes/exams.	3.71	1.09	Agree
to be in control of my learning in the course.	3.84	1.07	Agree
Overall,			
the video conferencing applications allowed me to express myself in new ways.	3.86	0.98	Agree
the video conferencing application was beneficial to my overall learning.	3.57	1.08	Agree

Academic institutions, teachers, and students relied on video conferencing applications to support virtual face-to-face classes being the substitute for students to communicate with their teachers (*M*=4.09; *SD*=0.93) and to attend classes remotely (*M*=4.16; *SD*=0.94). Students also agreed that the video conferencing applications helped them to feel a sense of community and social presence in the course (*M*=3.79; *SD*=1.09). Overall, using video conferencing applications allowed them to express themselves in new ways (*M*=3.86; *SD*=0.98) and were beneficial for their learning (*M*=3.57*; SD*=1.08).

The computed Pearson correlation coefficient presents a very weak relationship between the challenges encountered by the students and the student's perception of the use of video conferencing applications during synchronous learning. With 0.05 level of significance, the obtained p value=0.096, therefore accepting the null hypothesis that there is no significant relationship between the challenges encountered by the students during synchronous learning and the students' perception of the use of the video conferencing application.

Students who participated in this study have stated that they have an available and stable internet connection at home, however, they often encounter lagging screen sharing and sometimes distorted audio and video quality. In addition, teachers are also allowing the students to record the synchronous learning to cater to those students who struggled from technical difficulties or those students under correspondence mode. As a solution, teachers are distributing reading materials, PowerPoint files, and other educational materials to look back whenever they experience technical difficulties. With these strategies, students can still follow and learn the course content which contributes to a very weak relationship between the technical difficulties and the students' assessment on the use of video conferencing applications.

3.4 Relationship between the Frequency of Use of the Features and Students' Perception on the Use of Video Conferencing Applications during Synchronous Learning

The computed Pearson correlation coefficient presents a moderately strong relationship between the frequency of use of the features and the students' perception on the use of video conferencing applications during synchronous learning. This may imply that when students have increased frequency of use, they will most likely be an active participant. With 0.05 level of significance, the obtained *p* value=0.00, therefore rejecting the null hypothesis that there is no significant relationship between the frequency of use of the features and the students' perception on the use of video conferencing applications.

Studies have shown that class conducted in synchronously has increased students' engagement brought by screen sharing feature and enhanced creativity when used with constructivist methods. Further, the use of video conferencing applications develops students' urge to collaborate with their classmates and teachers with the use of virtual whiteboards, group messaging, and shared notes following the instructions of their teachers in the synchronous learning (Samson, 2020; Zuo, Yang, Wang, & Lou, 2020). Therefore, student engagement is arguably vital to effective learning especially in synchronous learning using video conferencing application (Dixson, 2015).

4. Conclusion and Recommendation

This study aimed to describe the use of the video conferencing application facilitating behavioral engagement of the senior high school students of PUP Manila during synchronous learning of General Chemistry 1 and General Physics 1.

The result showed that teachers allow the opportunities for behavioral engagement to happen during synchronous learning by making all the features of the video conferencing applications available for the students to use. Hence, the use of a microphone and group messaging for students to participate in the learning discussion tends to be the most used feature to respond to the teacher's question or ask questions. Although, students encounter lagging screen sharing when teachers present their presentations or when playing videos from their files or from video-sharing websites during synchronous learning.

In terms of the student's perception of the use of video conferencing applications, it allowed them to attend classes remotely, communicate with their teachers, collaborate with their teachers, and learn the content of the course. Moreover, it helped them to present themselves in a new and in creative way as it is beneficial for their overall learning in the time of the pandemic.

However, the result showed a weak and very low significant relationship between the challenges encountered by the students and the students' perception on the use of video conferencing applications during synchronous learning notwithstanding the evidence of some studies presented in this study that technical issues and glitches in an online setup affect students' learning. Most likely, the teachers are providing alternative ways of delivering the learning content to students through distributing reading materials, recorded sessions, or educational videos to reduce the impacts of technical problems on students' learning and engagement.

Conversely, the result also presents a moderately strong relationship between the students' frequent use of the features of the video conferencing applications and the student's perception of the use of video conferencing applications during synchronous learning – when there is an increased frequency of use of these features, most likely students are active participants. It implies that students maximizing the use of the features of video conferencing application promotes active participation, engagement, and social learning.

However, this study was conducted nearly the end of the semester and not during the actual synchronous learning session of the students. Therefore, the result of the study is not generalizable and conclusive.

It is further recommended that the researcher conducts an observation in synchronous learning with video conferencing applications where a teacher delivers the learning content to the students. The researcher may ask the teacher to implement certain lessons and activities and describe how often do the students interact in the learning discussion to substantially gather credible evidence of behavioral engagement happening in an online setting.

Acknowledgements

The researcher expresses his gratitude to the Ateneo de Manila University for the assistance provided in furnishing this manuscript. The researcher also extends his sincerest thanks to the PUP admin for allowing him to conduct the study and to the teachers and students who participated in the study.

References

Abaci, S. & Goodrum, D. (2016) “Zoom @ IU: Evaluation report of the pilot implemented in 2015-2016”. Retrieved November 5, 2020 from http://next.iu.edu/reports

Al Mamun,A A., Lawrie, G., Wright, T. (2016) “Student Behavioral Engagement in Self-Paced Online Learning”. In S. Barker, S. Dawson, A. Pardo, & C. Colvin (Eds.), Show Me The Learning. Proceedings ASCILITE 2016 Adelaide (pp. 381-386).

Andrew, L., Maslin-Prothero, S., & Ewens, B. (2015). Enhancing the online learning experience using virtual interactive classrooms. Australian Journal of Advanced Nursing, 32(4), 22–31.

Armstrong, A., & Thornton, N. (2012). Incorporating Brookfield’s discussion techniques synchronously into asynchronous online courses. Quarterly Review of Distance Education, 13(1), 1–9.

Bernard, R. M., Abrami, P. C., Lou, Y., Borokhovski, E., Wade, A., Wozney, L., Wallet, P. A., Fiset, M., & Huang, B. (2004). How does distance education compare with classroom instruction? A meta-analysis of the empirical literature. Review of Educational Research, 74, 379-439.

Bott, E. (2020) "Zoom Alternatives: Best Video Conferencing Software for Business" Retrieved October 23, 2020, from https://www.zdnet.com/article/best-video-conferencing-software-and-services-for-business/

Department of Education (2020) “Official Statement on Enrollment Data” Department of Education, Retrieved October 20, 2020, from https://www.deped.gov.ph/2020/07/17/official-statement-11/.

Department of Education (2020) "DepEd: Assistance for Teachers, Learners a Priority amid COVID-19 situation" Department of Education, Retrieved October 20, 2020, from https://www.deped.gov.ph/2020/03/24/deped-assistance-for-teachers-learners-a-priority-amid-covid-19-situation/

DepEd Tambayan (2020) "DepEd recommends 3 platforms for video conferencing requirements" Retrieved November 6, 2020, from https://depedtambayan.org/deped-recommends-3-platforms-for-video-conferencing-requirements/.

Dixson, M.D. Measuring Student Engagement in the Online Course: The Online Student Engagement Scale (OSE). Online Learn. 2015, 19, 51–65.

Finol, M. (2020) "Asynchronous vs. Synchronous learning: A Quick Review" Retrieved November 11, 2020, from https://www.brynmawr.edu/blendedlearning/asynchronous-vs-synchronous-learning-quick-overview#:~:text=Synchronous%20learning%20refers%20to%20all,or%20smaller%20groups%20get%20together

Fredricks JA, Blumenfeld PC, Paris AH. School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*. 2004;74:59–109.

Hartwell, C. (2017) Engaging Students in a Synchronous Distance Setting: Asking Online Questions. *Journal on Empowering Teaching Excellence. 2017. Vol. 1, Issue 1.*

Hudson, T. M., Knight, V., & Collins, B. C. (2012). Perceived effectiveness of web conferencing software in the digital environment to deliver a graduate course in applied behavior analysis. Rural Special Education Quarterly, 31(2), 27–39

Masa, J., Diaz, L., Delgado, S., Esteban, P. (2014) Assessment of the Use of Synchronous Virtual Classrooms in Higher Education. *The New Educational Review*. 38. 223-237

May, T. (2020) "7 best video conferencing tools of 2020" Retrieved October 23, 2020, from https://www.creativebloq.com/features/video-conferencing-tools

Moser, S. & Smith, P. (2015) "Benefits of Synchronous Online Courses" 2015 ASCUE Proceedings (pp. 43-48). Retrieved November 4, 2020, from https://files.eric.ed.gov/fulltext/ED571270.pdf

Park, Y. & Bonk, C. (2007) Synchronous Learning Experiences: Distance and Residential Learners’ Perspectives in a Blended Graduate Course. *Journal of Interactive Online learning.* Vol. 6 (3). ISSN: 1541-4914

Samson, P. (2020) Student behaviors in a blended synchronous course. *Journal of Geoscience Education*, 68:4, 324-333, DOI: 10.1080/10899995.2020.1768002

Zuo, M., Yan, Y., Wang, K., Luo, H. (2020). What Drives Rural Students’ Behavioral Engagement in Synchronous Online Classrooms? Examining the Effects of Discourse Interaction and Seating Location. 10.1007/978-3-030-51968-1_20.

The M in STEM and Issues of Data Literacy

Khalid KHAN & Jon MASON*
College of Indigenous Futures, Education, and Arts, Charles Darwin University, Australia
khalid.khan@cdu.edu.au
*jon.mason@cdu.edu.au

Abstract: This article examines post-pandemic scenarios that combine an emerging research agenda with three distinct topics relevant to the changing requirements of education systems worldwide: STEM education, mathematical thinking, and data literacy. In addition, how these topics are collectively positioned within the context of the digital environment is discussed. We examine the social usefulness of data literacy as may be embedded within a STEM paradigm and consider it as reaching beyond this paradigm into domains of civics and ethics. We propose that a meaningful strategy on 'data literacy' requires complex practical development and a curriculum presence explicitly within the Mathematics content area and across all STEM subjects, while also aligned to skills development typically associated with 21st century education together with emergent skills now associated with the fourth industrial revolution. In a similar way to how computational thinking has emerged as reaching beyond the formalities of computer science, the 'M' in STEM in this paper signifies 'mathematical thinking' is relevant beyond the discipline of Mathematics.

Keywords: STEM, inquiry, questioning, data literacy, curriculum, mathematical thinking

1. Introduction

In terms of calibrating teaching and learning to meet post-pandemic challenges, we question whether the core competencies of creativity, critical thinking, collaboration, and communication are sufficient to mitigate current and future societal challenges. Commonly known as the 4Cs that underpin many 21st century skills frameworks (World Economic Forum, 2016), such competencies have provided clear alignment with post-schooling employability requirements for the last two decades. But with characterizations of our times in terms of the 'misinformation age' (O'Connor & Weatherall, 2019), the 'black box society' (Pasquale, 2015) and 'surveillance capitalism' (Zuboff, 2019) the 'fourth industrial revolution' looks to be unleashing much more than technological innovation (Miller & Wendt, 2021). Even in the pre-pandemic world, data was easily sourced and routinely used to profile individuals and populations with the aim of predicting and modifying behaviors for financial gains (Zuboff, 2019; Couldry & Mejias, 2019). Post-pandemic, these practices will make more extensive use of student data. It is therefore timely to understand the challenges to traditional rights of individual privacy and social norms around behavioral modification that probably will be more technology directed from now. What does it mean to have agency and free will in a society with data as the key currency of transaction – and for some, seen as the 'new oil' (Humby, 2006)? There are several reasons why our societies should be concerned by the increasing dominance of data-driven epistemologies in social educational policies and marketing. One key reason is the diminishing control that students of all ages now have over how they are inscribed and profiled by 'smart' systems using such data and its influence of on their behaviors. How our on-line interactions, may potentially give rise to social justice and power issues due to data-driven decision-making, on education and governance and society. The question *who decides what for me?* has never been as prominent as it is now.

How can mathematical thinking and data literacy help? We know that computational thinking, design thinking, and systems thinking are all now embedded within STEM curricular, and that these modes of thinking are not enough in online educational environments where information and data tools can be easily manipulated. Within and beyond the STEM disciplines there are ethical and civic considerations. As Cowie and Cooper (2016) argue, 'every citizen needs to be data literate'. We also need to scaffold questioning techniques that probe deeper than simple Internet searching. A key issue this paper explores is how to focus STEM curricula on questioning techniques that shift traditional pedagogies from ‘thinking in answers’ to ‘thinking in questions’. Previously (Mason, Khan & Smith, 2020), we proposed a research question: *in what ways can we articulate a systemic approach to embedding data literacy within a STEM curriculum?* This paper represents a conceptual examination of aspects of mathematical thinking related to this question.

2. STEM and Data Literacy

In previous research we established that many definitions of data literacy are not fit for purpose since they don’t justify nor account for the changes data sets in recent times (Khan & Mason, 2016, 2017; Mason, Khan & Smith, 2019). As technology develops, an increasing number of common use objects are becoming ‘smart’. Various podiums silently collect our data. With innumerable children’s educational platforms in the post-pandemic world, safety and privacy issues will broaden in scope. This means that data literacy requires a better presence within school education. Without a clear framework and direction in curricula on data literacy and subjects where it needs embedding, this becomes challenging. Science now has moved to the ‘fourth paradigm’, where data-intensive investigations are becoming the norm (Tansley & Tolle, 2009). Little data that these days is collected with small gadgets like rulers, thermometers and stop watches has morphed into big data (Lohr, 2012, Mason et al., 2019). Data analysis is moving to investigate ‘big problems’ of the society through patterns in data.

Because data is produced and collected in diverse ways across all disciplines particularly under STEM areas, a systemic way of dealing with it across the curriculum is required. However, an inquiry problem in STEM often proceeds as: understanding the scope of the question or problem; data generated, analyzed, results interpreted and visualized, conclusion; more precise measurements, more variables, more data and the cycle repeats. (Mason, Khan & Smith, 2019).

In connecting STEM and Data Literacy, Cook and Bush (2015) used an integrated approach across their science and mathematics methods undergraduate courses to “support [pre-service teachers’] understanding of how to analyze and interpret data and their ability to teach it in their future classrooms” (p. 31). While not foregrounding the STEM acronym in the Australian Curriculum, the modes of thinking underpinning it (design thinking, systems thinking, and computational thinking) are presented as working together within the Technologies learning area.

The current Australian Curriculum touches upon on issues concerning data usage, though ‘data production’ (‘upstream’ data) is not yet explicit, nor is any provenance dealt with in terms of the origins and destinations of data. Given that data can be produced automatically and intentionally within digital environments, a good understanding of data literacy requires attention to the full scope of data production.

3. Methodology

In the current study, we have analyzed recent reviews and proposed recommendations for the Australian Curriculum *Foundations to Year 10*. Specific focus included inquiry and reasoning proficiencies within the Science, Technology and Mathematics learning areas. These were considered from a STEM perspective as an integrated problem-solving skill for innovative solutions. *Data Literacy* is the common thread that connects inquiry and questioning-based pedagogical practices in STEM.

4. The Australian Curriculum Review

In early 2021, the Australian Curriculum, Assessment and Reporting Authority (ACARA) initiated a public consultation to support its third review of the Australian Curriculum (AC). Two of the key recommendations include “mathematics classes need to have more units on financial literacy… and that primary school students are taught to have a greater awareness of online security” (Davies, 2021). However, focus has remained mostly on semantics – for example, during the public consultation of the AC review, ACARA has proposed renaming ‘ICT Capability’ as ‘Digital Literacy’.

Obviously, curriculum review requires more in-depth analysis when connecting Mathematics and Science Curriculums with data literacy and computational thinking that digital literacies require.

4.1 Australian Science Curriculum

The Australian Science curriculum consists of three strands – Science Understanding, Science as a Human Endeavour, and Science Inquiry Skills. The inquiry strand is presented as: 'Identifying and posing questions; Planning, conducting and reflecting on investigations; Processing, analyzing and interpreting evidence; and Communicating findings. This strand is concerned with evaluating claims, investigating ideas, solving problems, drawing valid conclusions *and developing evidence-based arguments.'* (ACARA, 2012)

The proposed revision of science curricula puts some focus on the role of data by including analysis of diverse data and information to identify and explain patterns, trends, relationships and anomalies (ACRC, 2021). Inquiry skills, as constructivist pedagogy requirements, are a key feature of the Australian Curriculum in Science, History and Geography. Inquiry learning considers teachers and students as co-learners and co-constructors of knowledge (Callison, 2006). Apart from developing questioning skills, inquiry processes include collecting and analyzing data and information for higher order problem solving. These include critical thinking, reasoning, and reflecting. Issues concerning data literacy are not foregrounded in questioning or examining data to identify ethical considerations, biases, personal and private data, etc. In other words, issues related to data literacy aren't explicit and do not (yet) have a clear presence in the curriculum.

4.2 Australian Mathematics Curriculum

The Australian Mathematics Curriculum consists of four 'general proficiencies' – Understanding, Fluency, Problem Solving, and Reasoning. The proposed revision within the Australian Curriculum suggests removing outdated and non-essential content and replacing it with content that is more contemporary. The critical processes of reasoning and problem-solving have been suggested to have more identifiable presence within content and achievement standards now. The proposed revision also gives teachers better clarity and guidance about what they are expected to teach. (ACRC, 2021).

Data Literacy has an explicit presence within Mathematics. Some (e.g., Gould, 2017) argue Data Literacy is essentially equivalent to Statistical Literacy. Data Literacy makes a clear presence under Statistics and Probability. Overall, the basic skills that students (and teachers both) need for improving decision making skills using data are: (1) Knowledge of diverse data collection protocols; (2) Selecting protocols best suited to answer (students' and teachers') questions; (3) Collating and graphing data; (4) Discerning trends and differences in data; (5) Using data in team problem solving; and (6) Selecting evidence-based interventions The following tables provide a comparison of actions that Australian Curriculum Science Inquiry Skills and Problem Solving and Reasoning in Mathematics:

Table 1. *Science Inquiry Skills (adapted from ACARA, 2012)*

Science Inquiry Skills	Attributes
Questioning and Predicting	Identify, chose, select, pose, formulate Questions (everyday life contexts). Identify and investigate data in familiar and unfamiliar contexts. Use previous investigations and existing data. Hypothesize, predict, revise and refine questions. Think" what will happen if…". Take into consideration the social, cultural, economic, environmental or moral aspects.
Planning and Conducting Investigations	Gather, explore, sort, classify information. Manipulate objects and make observations. Research ideas collaboratively. Consult. Explore multiple ways to collect and record data. Use informal measurements. Explore multiple ways to approach a problem. Be fair. Accuracy vs Approximations. Consider safe processes. Use primary and secondary sources. Understand fair investigation methods. Know that equipment may influence the reliability of collected data and results. Model and Simulate.
Processing and analyzing data and information	Use drawings to represent observations. Use a range of methods to sort data. Identify similar, odd-one-outs and opposites. Group items using similarities and differences. Analyze. Compare predictions with results. Identify patterns/trends. Suggest reasons for the findings. Use a range of representations. Construct tables, graphs. Use digital organizers of data such

	as, spreadsheets. Identify data that support/ negate the hypothesis. Understand there could be more than one possible explanation for results. Review scientific understanding.
Evaluating and Reflecting	Compare and evaluate observations. Discuss similarities/ differences. Consider indicators of quality of the data. Describe experiences. Reflect whether a test was fair? Research methods used by scientists. Suggest improvements. Evaluate conclusions. Identify sources of uncertainty & possible alternative explanations. Identify gaps or weaknesses. Identify alternative explanations consistent with data. Critically analyze the validity of information. Describe how scientific and data driven arguments can be used to make decisions.
Communicating	Represent and communicate in a variety of ways. Use formal and informal representations. Acknowledge and explore 'other' ways of communication. Label diagrams. Present research and results using other forms of representation of data and scientific language appropriate for the target audience. Use secondary sources and students' own findings to help explain a scientific concept. Use internet and on-line data to facilitate collaboration and discussion.

Table 2. *Problem Solving & Reasoning strands in Mathematics (adapted from ACARA, 2012)*

Reasoning	Develop capacity for logical thought and actions. Analyze, prove, evaluate, explain, infer, justify and generalize. Deduce /justify strategies and conclusions reached. Adapt the known to the unknown. Transfer learning from one context to another. Prove something is true or false. Compare related ideas and explain choices.
Problem-Solving	Make choices, interpret, formulate, model and investigate problem situations. Communicate solutions effectively. Use mathematics to represent unfamiliar or meaningful situations. Design investigations and plan approaches. Apply existing strategies to seek solutions. Verify reasonableness of their answers.

One of the key recommendations under the revised Australian Curriculum is for year 10 students to critically analyze media in terms of the claims and conclusions, noting limitations and potential sources of bias. Table 3 is constructed for Data Literacy attributes:

Table 3. *Data Literacy under Statistics and Probability F-10 (adapted from ACARA (2012)*

Mathematical Thinking	Attributes
Representation and Interpretation	Collect information, make inferences. Question. Collect categorical/numerical data. Determine questions. Identify categories. Represent data with/without digital technologies. Compare various representations, describe similarities and differences. Describe/interpret different data sets. Identify questions. Identify data sources. Plan methods of data collection and recording data. Construct, interpret and compare a range of data displays. Summarize data by calculating measures of center and spread. Make sense of the data. Construct stem-and-leaf plots and histograms. Use these to compare two like sets of data. Describe the shape of the distribution. Understand key terms. Describe and interpret data sets using location (center) and spread. Compare means, medians and ranges of two sets of data.
Uncertainty	Identify and describe chance events. Understand and describe outcomes. Assign probabilities. Identify variations. List outcomes. Justify. Investigate. Construct sample spaces. Conduct repeated trials. Investigate probabilities. Compare experiments which differ. Calculate relative frequencies. Use Venn diagrams & two-way tables. Investigate reports in digital media (and elsewhere) use data to estimate population means and medians. Analyze claims, inferences and identify ethical considerations.

4.3 Australian Technology Curriculum

The Technologies curriculum under ACARA aims to provide students opportunities to consider how solutions with future perspectives. It asks to identify benefits and risks and weigh impacts using critical and creative thinking (ACARA, 2012). Data Literacy is implicit as part of Computational Thinking (CT) within the Digital Technologies Curriculum, According to ACARA Computational thinking (CT) is:

> *a problem-solving method that is applied to create solutions that can be implemented using digital technologies. It involves integrating strategies, such as organising data logically, breaking down problems into parts, interpreting patterns and models and designing and implementing algorithms.*

> *Computational thinking is used when specifying and implementing algorithmic solutions to problems in Digital Technologies. For a computer to be able to process data through a series of logical and ordered steps, students must be able to take an abstract idea and break it down into defined, simple tasks that produce an outcome. This may include analysing trends in data, responding to user input under certain preconditions or predicting the outcome of a simulation.*
>
> *This type of thinking is used in Design and Technologies during different phases of a design process when computation is needed to quantify data and solve problems.* (ACARA, 2012).

The review requires meaningful connections with Mathematics through data representation (ACRC, 2021). With the current expectations changing, it is proposed students should be able to "*possess and be able to demonstrate computational thinking skills that include pattern recognition, decomposition, determining which (if any) computing tools could be employed in analyzing or solving the problem, and defining algorithms as part of a detailed solution*." (OECD 2018, p. 7). PISA2021 also points out that "*Long-term trajectory of mathematical literacy should also encompass the synergetic and reciprocal relationship between mathematical thinking and computational thinking*" (OECD 2018, p. 7). Extending these ideas, Ho (2021) explored Computational Thinking through the pedagogy of Mathematics and highlighted the 'data principle' as one of the four core principles for task design – which requires the educator to identify if the topic manifests instances and common traits/trends/patterns that can be observed, quantified, stored and treated as data?

Within the machine learning paradigm, algorithms learn from and recognize patterns (within the data) and when new data is received as an input, recognize, distinguish and identify the same patterns to define and match answers. This process is same as how a child learns from environmental contexts and in a mathematics class, e.g., identifying and naming. Figure 2 depicts how Data (and hence Data Literacy) is pivotal Computational Thinking, Mathematical Thinking and Scientific Inquiry.

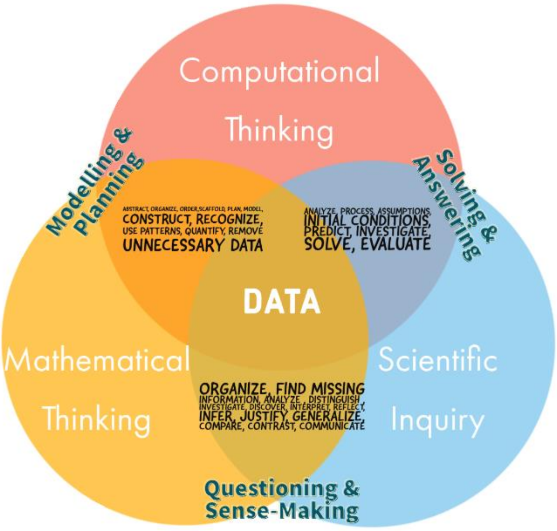

Computational Thinking
Modelling & Planning
ABSTRACT, ORGANIZE, ORDER,SCAFFOLD, PLAN, MODEL, CONSTRUCT, RECOGNIZE, USE PATTERNS, QUANTIFY, REMOVE UNNECESSARY DATA
Solving & Answering
ANALYZE, PROCESS, ASSUMPTIONS, INITIAL CONDITIONS, PREDICT, INVESTIGATE, SOLVE, EVALUATE
DATA
Mathematical Thinking
Scientific Inquiry
ORGANIZE, FIND MISSING INFORMATION, ANALYZE , DISTINGUISH, INVESTIGATE, DISCOVER, INTERPRET, REFLECT, INFER, JUSTIFY, GENERALIZE, COMPARE, CONTRAST, COMMUNICATE
Questioning & Sense-Making

Figure 2. Data at the Intersection of STEM.

STEM is not just focusing on four key learning areas, but also designing learning and an interdisciplinary approach to problem solving that draws on deepening and comprehensive understanding of Mathematics, Science, Engineering and Technology. Salmacia (2017) points out that '...data literacy skills were some of the most important that a teacher could possess [... and that] becoming an outcome-driven, data-literate teacher was absolutely necessary [...]' (p. 140) Additionally, he concludes that addressing culturally responsive practices related to assessments and other data collection instruments, and as well as providing technology instruction with a view on the key role that technology and computers play with in data literacy paradigm (e.g., skills related to the use of data dashboards, data systems, spreadsheet functions, etc.) are critical.

5. Conclusion

While inquiry-based approaches are common to most subject areas across educational curricula, they are also shaped by the specific learning areas and in practice. Pedagogical approaches will vary, and no one approach whether student-centred or teacher-directed in emphasis is appropriate for all contexts. Understanding the scope of data and its implications in dealing with current and future challenges is a key theme in this paper. This requires that educational curricula and content are routinely reviewed and updated. The whole notion of '21st century skills' is thus becoming a questionable construct. Requirements for inquiry and reasoning using digital platforms are becoming more complex in nature requiring Data Literacy skills within Mathematics and STEM curriculum areas. Given the pace of change, we suspect that such a shift will take several iterations.

References

Akinshin, A. (2021). Misleading histograms | Andrey Akinshin. Retrieved 14 May 2021, from https://aakinshin.net/posts/misleading-histograms/.

Australian Curriculum, Assessment and Reporting Authority (ACARA), (2021). The Australian Curriculum Review. Public consultation https://www.australiancurriculum.edu.au/consultation/mathematics/

Australian Curriculum Review consultation (ACRC), (2021). Retrieved 24 May 2021, from https://www.australiancurriculum.edu.au/consultation/

Callison, D., & Preddy, L. (2006). *The blue book on information age inquiry, instruction, and literacy*. Libraries Unlimited Incorporated.

Conn, C., Boham, K., Hernandez, N., Powell, P., & Luzader, J. (2020). Teaching and Assessing Data Literacy: Resource Guide for Supporting Pre-Service and In-Service Teachers. Ceedar. Education. https://ceedar.education.ufl.edu/wp-content/uploads/2020/10/Arizona_Data_Literacy_Resource_Guide.pdf

Cook, K. L., & Bush, S. B. (2015). Structuring a science mathematics partnership to support preservice teachers' data analysis and interpretation skills. *Journal of College Science Teaching, 44*(5), 31-37

Couldry, N. & Mejias, U. A. (2019). *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating it for Capital*. Stanford, CA: Stanford University Press.

Cowie, B., & Cooper, B. (2017). Exploring the challenge of developing student teacher data literacy. *Assessment in Education: Principles, Policy & Practice, 24*(2), 147–163.

Davies, L. (2021). *Curriculum proposals deserve support, not another iteration of the history wars*. The Sydney Morning Herald. Retrieved 21 May 2021, from https://www.smh.com.au/education/curriculum-proposals-deserve-support-not-another-iteration-of-the-history-wars-20210502-p57o7a.html

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal, 16*(1), 22-25.

Ho, W.K. (2021). *Computational Thinking Through the Lens of a Mathematics Educator*. CTE-STEM 2021 Keynotes. https://cte-stem2021.nie.edu.sg/keynote.html

Humby, C. (2006), quoted in Mavuduru, A. (2020). Is Data really the New Oil in the 21st Century? Towards Data Science. https://towardsdatascience.com/is-data-really-the-new-oil-in-the-21st-century-17d014811b88

Khan, K. & Mason, J. (2019). Examining the Data to Identify Essential Questions – Guilty before Innocent. *International Journal of Smart Technology and Learning. 1*, 3, p. 244-266 23 p.

Khan, K. & Mason, J. (2018). Faking it with Data – A Curriculum Perspective. Proceedings, 'Developing Real-Life Learning Experiences: Using Innovation and Enhanced Technologies', King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand. June 4-10.

Khan, K. & Mason, J. (2017). Learning to be Data Smart. *Workshop Proceedings of the 25th International Conference on Computers in Education*. New Zealand: Asia-Pacific Society for Computers in Education.

Mason, J., Khan, K. & Smith, G. (2020). A SySTEMic approach to Data Literacy. In Chang, M., So, H-J., Wong, L-H., et. al., (eds.). *Proceedings of the 27th International Conference on Computers in Education*. Asia-Pacific Society for Computers in Education, 665-667.

Miller, K., & Wendt, K. (Eds.). (2021). *The Fourth Industrial Revolution and Its Impact on Ethics: Solving the Challenges of the Agenda 2030*. Springer Nature.

O'Connor, C., & Weatherall, J. (2019). *The misinformation age: How false beliefs spread*. Yale University Press.

OECD. (2018). PISA 2021 Mathematics Framework (Draft). November 2018.

Pasquale, F. (2015). *The Black Box Society*. Harvard University Press.

Salmacia, K. A. (2017). *Developing outcome-driven, data-literate teachers*. Doctoral dissertation, University of Pennsylvania.

World Economic Forum. (2016). *New Vision for Education: Fostering Social and Emotional Learning through Technology* (p. 36). http://www3.weforum.org/docs/WEF_New_Vision_for_Education.pdf

Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York, NY: Public Affairs

Development and Preliminary Evaluation of the Learning Potential of an Online System in Support of a Student-Generated Testlets Learning Activity

Fu-Yun YU
Institute of Education, National Cheng Kung University, Taiwan
fuyun.ncku@gmail.com

Abstract: Currently, around fifty online learning systems have been developed to support student-generated questions (SGQ) activities, highlighting the various affordances of computers and networked technologies. While a variety of question types are commonly supported in these online SGQ systems, no systems are in existence that support student generation of the testlet format — an attractive test format that is becoming increasingly prevalent and prominent in contemporary testing and assessment practices. In view of this research gap, this study is aimed toward two research goals: first, to develop an online learning system in support of student-generated testlets learning activities, and second, to preliminarily assess its learning potential. Two general design principles that guide the development of the student-generated testlets online learning system (i.e., flexibility and cognitive support) are described first. Afterwards, the evaluation study used to assess the learning potential of student-generated testlets as compared to SGQ is presented. Based on the data collected from a sixth-grade class of students (n = 28), it was found that significantly more participants felt the online testlet-generation task as better promoting learning as compared to SGQ, $X^2 = 9.93$, $p < .01$. Furthermore, based on the result of the constant comparative method done on the participants' provided explanatory reasons, the distinct feature of student-generated testlets, specifically, the situation/scenario, and its possible positive and negative effects on learning were highlighted to provide suggestions for future studies.

Keywords: Design principles, development of online learning systems, evaluation study, student-generated questions, student-generated testlets

1. Introduction

Characterized by students generating questions and their corresponding answers (explanations) in response to an exposed instructional event (e.g., a given problem, a focal topic, reading material, a learning activity, and so on) for learning and assessment purposes (Yu, Wu, Hung, 2014), student-generated questions (hereinafter, SQG) is regarded as a cognitively and metacognitively engaging learning activity (Rosenshine, Meister, & Chapman, 1996). Empirical evidence on the effects of SGQ on the promotion of student cognitive and affective effects have generally been positive (Rosenshine et al., 1996; Rosli, Capraro, & Capraro, 2014; Zuya, 2017).

Since the turn of the twentieth century, in light of the distinctive affordances of computers and networked technologies (e.g., time- and place-independence, immense data storage capability, fast processing speed, multi-mode interaction, and multimedia representation, among others) (Yu, 2009; Yu & Su, 2015), around fifty online learning systems in support of SGQ activities have been developed around the world. With question-generation as the core function, existing online SGQ systems usually support the generation of different question types (e.g., true/false, multiple-choice, short-answer, matching, fill-in-the-blank, etc.). However, to the best of the author's knowledge, currently, no systems support student generation of a testlet format.

In essence, testlet involves the formulation of a group of related question items on the basis of a given situation or scenario. Because of its efficiency in terms of item development and test administration, a testlet is an attractive test format among test developers (Keng, Ho, Chen & Dodd, 2008; Lane, Raymond & Haladyna, 2016), and its presence in contemporary testing and assessment practice is becoming increasingly prevalent and prominent.

At present, online systems supporting the formulation and sharing of student-generated testlets are not in existence, and the learning potential of student-generated testlets is yet to be realized. In view of these research gaps, in this study, the design and development of an online learning system in support of student-generated testlets is described, and its learning potential is evaluated by a preliminary study.

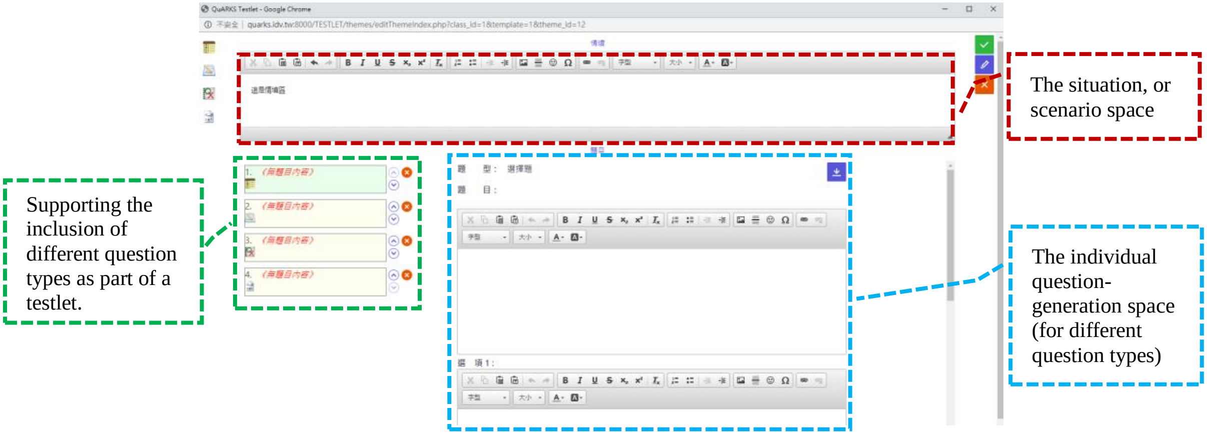
2. Design Principles Guiding the Development of an Online Student-Generated Testlets Learning System

To enable the developed system to be integrated for classroom use and to make it possible to support a versatile, scaffolded learning environment, design principles were set up in advance to guide its development. Explicitly, two general design principles were set up to guide the development of the student-generated testlets online learning system — flexibility and cognitive support. Each of the two principles are further broken down into sub-guidelines to help attain the system goals (i.e., creating a versatile, scaffolded learning space).

2.1 Flexibility in Testlet Generation

In terms of flexibility, three aspects of testlet-generation are highlighted: multiple-presentation of situations/scenarios, re-use and editability, and opportunity for self-expression.

Multiple-presentation of situations/scenarios (the top portion of Figure 1): Since the situations/scenarios in testlets in both text- and graphics-based formats (e.g., graphs, maps, diagrams, tables, charts, etc.) are commonly used in testing and assessment, the student-generated testlets online learning system allows situations/scenarios presented in either text or graphics form to support various stimuli deemed suitable by the testlet-author.


QuARKS Testlet - Google Chrome
不安全 | quarks.idv.tw:8000/TESTLET/themes/editThemeIndex.php?class_id=1&template=1&theme_id=12
情境
這是情境區
1. (無題目內容)
2. (無題目內容)
3. (無題目內容)
4. (無題目內容)
題　型：選擇題
題　目：
選　項 1：

The situation, or scenario space

Supporting the inclusion of different question types as part of a testlet.

The individual question-generation space (for different question types)

Figure 1. The Online Student-generated Testlets Learning System.

Re-use and editability: To reflect and actualize the core concepts and value of Web 2.0 (essentially, innovation in assembly and the remixing of data) (Anderson, 2007), rather than limiting the learners to generating a set of question items based on a given, fixed situation/scenario, the system allows the learners to (a) edit a given situation/scenario (by deleting, re-arranging, or adding the conditions given within) before proceeding to the task of generating a set of question items or to (b) work from a given situation/scenario with a set of related question items. For this purpose, the system supports transfer of a set of testlets from different sources (i.e., peers and the teacher) as a start or reference point for student-generated testlets.

Opportunity for self-expression: To enable the creative writing opportunity to be better actualized, the system allows learners to create their own situations/scenarios (before proceeding to generating the corresponding question items set) rather than working from a given situation/scenario. With this, the system supports a more open, less constrained testlet generation learning space that is promotive of self-expression.

2.2 Cognitive Support for Testlet Generation

To help students generate meaningful, relevant situations/scenarios and the corresponding interconnected set of question items, cognitive support mechanisms are envisioned. Specifically, two types of cognitive support are embedded in the system — conceptual framework for scenarios and item set generation and context-dependent examples.

For the conceptual framework for the scenarios and item set generation, a set of existing frameworks, such as the story grammar category (Nolte & Singer, 1985; Knudson, 1988) and main ideas schemes (Ritchie, 1985) is provided to guide the creation of scenarios whereas the "what if" strategy (Brown & Walter, 2005) is offered for editing given scenarios and question items sets.

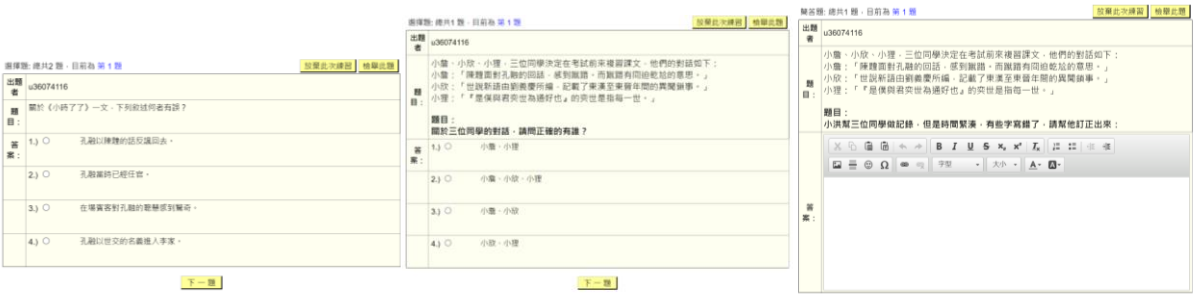
At the same time, spaces for the provision of context-dependent examples by individual instructors to assist student-generated testlet activities are also built in.

3. The Preliminary Study Assessing the Learning Potential of the Developed Online Student-Generated Testlets Learning System

The learning potential of the developed online student-generated testlets learning system is the focus of this evaluative study. In particular, student perceptions of the comparative learning usefulness of student-generated testlets and student-generated questions (SGQ) are examined.

3.1 Methods

Students from a sixth-grade class (*N* = 28) in a single primary school in the southern part of Taiwan participated in this preliminary evaluation study for eight consecutive weeks. As a routine, each week after attending five 40-minute instructional sessions on Chinese, the participants used iPads to access the developed online learning systems for practice sessions on the Chinese lesson covered in the current week during their weekly 40-minute alternative curriculum. For the weekly practice session, the participants first engaged in question- and testlet-answering activities on the currently learned Chinese lesson (see Figure 2) before moving on to question- or testlet-generation activities.



選擇題: 總共2 題，目前為 第 1 題

出題者 u36074116

題目：關於《小時了了》一文，下列敘述何者有誤？

答案：
1.) ○ 孔融以陳韙的話反諷回去。
2.) ○ 孔融當時已經任官。
3.) ○ 在場賓客對孔融的聰慧感到驚奇。
4.) ○ 孔融以世交的名義進入李家。

選擇題: 總共1 題，目前為 第 1 題

出題者 u36074116

題目：
小詹、小欣、小理，三位同學決定在考試前來複習課文，他們的對話如下：
小詹：「陳韙面對孔融的回話，感到踧踖，而踧踖有侷促尷尬的意思。」
小欣：「世說新語由劉義慶所編，記載了東漢至東晉年間的異聞瑣事。」
小理：「『是僕與君奕世為通好也』的奕世是指每一世。」

題目：
關於三位同學的對話，請問正確的有誰？

答案：
1.) ○ 小詹、小理
2.) ○ 小詹、小欣、小理
3.) ○ 小詹、小欣
4.) ○ 小欣、小理

簡答題: 總共1 題，目前為 第 1 題

出題者 u36074116

題目：
小詹、小欣、小理，三位同學決定在考試前來複習課文，他們的對話如下：
小詹：「陳韙面對孔融的回話，感到踧踖，而踧踖有侷促尷尬的意思。」
小欣：「世說新語由劉義慶所編，記載了東漢至東晉年間的異聞瑣事。」
小理：「『是僕與君奕世為通好也』的奕世是指每一世。」

題目：
小洪幫三位同學做記錄，但是時間緊湊，有些字寫錯了，請幫他訂正出來：

答案：

Figure 2. Question-answering (left) and Testlet-answering (Middle and Right).

To enable the participants to make comparisons of question- and testlet-generation, this evaluation study consisted of two main phases: online question/testlet-answering activities with question-generation (Phase I) and online question/testlet-answering activities with testlet-generation (Phase II). Two online learning systems were adopted to support online drill-and-practice, question-generation, and testlet-generation activities — QuARKS for the first two activities (Yu, 2009) and the newly developed online system for the last one.

At the beginning of Phases I and II, orientation and training sessions on essential knowledge and skills to ensure meaningful engagement and successful completion of the integrated learning task were arranged. Specifically, during Phase I orientation and training session, question-generation techniques, including main ideas, target words proposed by Yu and Pan (2014), and what if/what if not proposed by Brown and Walter (2005) were explained after introduction to the online drill-and-practice function in QuARKS. In addition, the criteria for high-quality student-generated questions and the operational procedures involved for question-generation in QuARKS were explicated before the participants proceeded to question-generation activities. For each question-generation activity, at least three questions of any of the three question types of the participants' choice were suggested (i.e., true-false, multiple-choice, or short-answer) (see Figure 3).

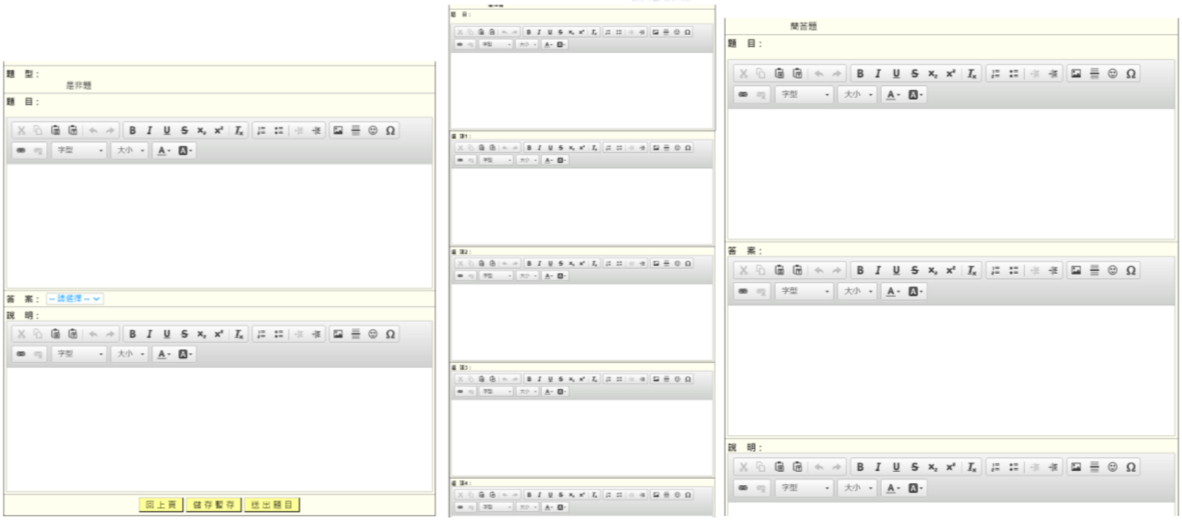


Figure 3. Generation of Different Question Types in Quarks in the Preliminary Evaluation Study: true/false (left), multiple-choice (middle), and short-answer (right).

As for Phase II orientation and training session, techniques and criteria associated with testlet-generation and the navigating procedures within the newly developed system were described. For each testlet-generation activity, the participants were directed to generate at least one testlet consisting of the scenario and at least two question items that were either true-false, multiple-choice, or short-answer (see Figure 1).

After the conclusion of the last session, one closed-ended question with explanations for their selection was distributed to collect the participants' thoughts on the relative learning usefulness of online SGQ and testlet-generation — which of the two activities do you think better help your learning of Chinese: online question-generation, online testlet-generation, two at about the same level. Please explain your selection.

3.2 Results and Discussion

As shown in Table 1, more than half of the participants perceived 'online testlet-generation' as better promoting their learning of the focal subject matter (i.e., Chinese). Comparatively, significantly fewer participants voted for 'online question-generation,' and one-fourth of the participants felt that the two activities were at about the same level in terms of learning usefulness. Furthermore, an X^2 test on the observed frequency distribution among the three options conducted was statistically significant, X^2 = 9.93, p < .01.

Table 1. *Descriptive and Inferential Statistics of Perceived Learning Usefulness of Online Question-Generation and Online Testlet-Generation* (N = 28)

	Online question-generation *f* (%)	Online testlet-generation *f* (%)	About the same level *f* (%)	X^2	*p*
Learning usefulness	4 (14.29%)	17 (60.71%)	7 (25%)	9.93	.007

The constant comparative method proposed by Lincoln and Guba (1985) was adopted to analyze the descriptive explanations provided by the participants for their selections. The results in respect to the three options were presented and discussed separately.

First, for those four rooting for online question-generation, one major theme emerged. The four explanatory responses all pointed to the '*less restrictive*' nature of question-generation. As the participants succinctly explained, '*without the restriction of the given situation/scenario, more* *questions with a different focus can be generated*' and '*time and effort can be invested and focused on the main ideas covered in the Chinese lesson*,' not on '*writing up or revising the given scenario*.'

As for those supporting online testlet-generation, two major themes emerged. First, five out of the seventeen participants pointed out the fact that there are essentially two tasks involved in testlet-generation, namely scenario- and question-writing, which '*naturally led to deeper, better, and more learning*.' Second, another five in this group of participants pointed to the idea that the given *situation/*scenario served as an anchoring point and helped '*set a premise*' or '*reference point*.' Of these, two respondents further added that the situation/scenario made the task '*more difficult, cognitively demanding, yet helped lead to deeper and better learning*.' Alternatively, two others felt the situation/scenario made the task '*easier, as it helped direct attention*' and led to '*the learning of the important aspects of the material*.' Disregarding whether the situation/scenario made the task harder or easier as the respondents perceived it, it reflected some ideas related to anchored instruction proposed by The Cognition and Technology Group at Vanderbilt University (1990). Anchored instruction emphasizes tasks intended to engage learners (i.e., the question-generation task in this study, whereas it is the problem-solving task in anchored instruction) would base on or tie around a presented scenario, which serves as an anchor to help build meaningful connections with the instructional content (The Cognition and Technology Group at Vanderbilt, 1990).

Finally, for those who felt question- and testlet-generation provided a similar level of learning usefulness ($N = 7$), despite the fact that three noted the differences in task requirements between these two activities (i.e., testlets having the extra given situation/scenario to work on), the thought that both activities '*essentially dealt with the core act of question-generation*' such that both helped Chinese learning to about the same degree was commonly shared by this group of respondents.

4. Conclusions

In this study, an online learning system in support of testlet-generation was developed, following a set of pre-set design guidelines. A preliminary study was conducted to evaluate its learning potential, as compared to the increasingly accepted and empirically proven SGQ. As revealed in this study, significantly more participants rooted for the online testlet-generation task as better promoting learning. Despite this, interesting phenomenon was revealed from the explanatory responses provided by the participants. While some valued the situation/scenario given in testlets for serving as an anchoring point and help '*set a premise*' or '*reference point*,' others felt strongly about '*the restrictions the scenario* may impose on' which may inadvertently affect the fluency of the learning and thinking process. With such contrasting perceptions towards given situations/scenarios, which is one essential, distinct component of testlets, and the preliminary nature of this evaluation study, issues regarding if there are differential learning effects between online testlet-generation and question-generation would be a topic worthy of examination via a more stringent experimental research method. Specifically, with the insights gained from the responses provided by the participants, learning effects on perceived task value, perceived task difficulty, learning motivation, task performance, and academic performance would be worthwhile areas to be targeted in future research.

Acknowledgements

This study was supported by the Ministry of Science and Technology in Taiwan (MOST 108-2511-H-006-007-MY3). Thanks are extended to Ya-Hsin Chang who served as the research assistant to assist in collecting the evaluative data reported in this work.

References

Anderson, P. (2007). What is Web 2.0? Ideas, technologies and implications for education. *JISC Technology and Standards Watch*, Feb. 2007. Retrieved May 30, 2021 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.108.9995&rep=rep1&type=pdf

Brown, S. I., & Walter, M. I. (2005). *The art of problem posing* (3rd ed.). New Jersey: Lawrence Erlbaum Associates.

Keng, L., Ho, T-H., Chen, T-A. & Dodd, B. (2008). A comparison of item and testlet selection procedures in computerized adaptive testing. Paper presented at the *National Council on Measurement in Education*. March 24th, New York City.

Kopparla, M., Bicer, A., Vela, K., Lee, Y., Bevan, D., Kwon, H., ... & Capraro, R. M. (2019). The effects of problem-posing intervention types on elementary students' problem-solving. *Educational Studies, 45*(6), 708-725.

Knudson, R. E. (1988). The effects of highly structured versus less structured lessons on student writing. *Journal of Educational Research, 81*, 365-368.

Lane, S., Raymond, M. R. & Haladyna, T. M. (2016). *Handbook of test development* (2nd Ed). NY: Routledge.

Lincoln, Y. S. & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage Publications.

Moses, B. M., Bjork, E., & Goldenberg, E. P. (1993). Beyond problem solving: Problem posing. In Brown S. I. & Walter M. I. (Eds.). *Problem posing: Reflections and applications* (pp. 178–188). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nolte, R. Y., &Singer, H. (1985). Active comprehension: Teaching a process of reading comprehension and its effects on reading achievement. *The Reading Teacher, 39*, 24-31.

Ritchie, P. (1985). The effects of instruction in main idea and question generation. *Reading Canada Lecture, 3*, 139-146.

Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: A review of the intervention studies. *Review of educational research, 66*(2), 181-221.

Rosli, R., Capraro, M. M., & Capraro, R. M. (2014). The effects of problem posing on student mathematical learning: A meta-analysis. International Education Studies, 7(13), 227-241.

The Cognition and Technology Group at Vanderbilt (1990). Anchored instruction and its relationship to situated cognition. *Educational Researcher. 19*(6), 2–10.

Vreman-de Olde, C., & de Jong, T. (2006). Scaffolding learners in designing investigation assignments for a computer simulation. *Journal of Computer Assisted Learning, 22*(1), 63-73.

Yu, F.-Y. (2009). Scaffolding student-generated questions: Design and development of a customizable online learning system. *Computers in Human Behavior, 25*(5), 1129-1138.

Yu, F. Y. & Pan, K-J (2014). Effects of student question-generation with online prompts on learning. *Educational Technology and Society, 17*(3), 267-279.

Yu, F.-Y., & Su, C.-L. (2015). A student-constructed test learning system: The design, development and evaluation of its pedagogical potential. *Australasian Journal of Educational Technology, 31*(6), 685-698.

Yu, F. Y., Wu, C. P., & Hung, C-C (2014). Are there any joint effects of online student question generation and cooperative learning? *The Asia-Pacific Education Researcher, 23*(3), 367-378.

Zuya, H. E. (2017). The benefits of problem posing in the learning of mathematics: A systematic review. *International Journal of Advanced Research, 5*(3), 853-860.

Learn to Design (L2D): a TPD Program to Support Teachers in Adapting ICT Learning Materials to Their Local Context through Research-Based Strategies

Gaurav JAISWAL[a*], Sunita RASTE[*]& Sahana MURTHY[*]
[a]*Indian Institute of Technology Bombay, India*
*193380003@iitb.ac.in

Abstract: Knowledge of ICT and ability to integrate technology effectively in one's teaching practice has gained enormous importance for today's teachers. In the last two decades a lot of quality teaching materials and repositories have been created worldwide. However, capacity building of teachers and improving teachers' learning design skills has not received as much attention. Hence teachers are not adequately prepared to use technology in their practice and integrate it with learner-centric pedagogical strategies. While a large amount of resources and teaching and learning materials are available for common use, these lack contextualization. There is a need for teachers to build the necessary knowledge and skills to adapt the materials for their context. In this paper we describe the design and implementation of a TPD program, *Learn to Design* for supporting teachers to design research-based strategies using learning materials available on the DIKSHA platform, a national initiative in India, and adapt them to their local context. The program design draws from the TPACK framework and includes hands-on activities focused on enhancing the ability and self-efficacy of teachers to integrate technology into their teaching. The TPD workshop has been implemented with ~300 participants who were in-service teachers from across different states in India. The study analyzes data from multiple sources. The findings show improvement in participants' self-perception and ability to effectively integrate technology, as well as high intention to implement such strategies in their own classroom. In *Learn to Design*, we have created the design and session plan to support teachers in effectively integrating content, pedagogy and technology in their own context.

Keywords: Teacher professional development, contextualization of learning materials, ICT integration, TPACK, Constructive alignment

1. Introduction

The affordances of information and communication technologies in teaching and learning provides a number of benefits and should be used to support student-centered learning. (Howland, Jonassen & Marra, 2012). However, inadequate teacher preparation to use technology and apply new educational strategies (Brown & Warschauer, 2006), as well as teachers' ideas and attitudes toward technology (Ertmer, 2005), have hampered meaningful integration of technology with successful pedagogical practices. Another issue to examine is the teacher's own view of the need for new technology training (Amhag, 2019). Teachers must not only have digital competence for their own professional practice, but they must also act as role models for their students when using digital technology in the classroom (Lund and Erikson 2016). However, achieving acceptable levels of digital proficiency is not without difficulties, (Gudmundsdottir and Hatlevik 2018). There are 3 types of barriers to effective ICT implementation - extrinsic, intrinsic factors and design thinking (Ertmer, 1999) . Researchers have argued that overcoming the extrinsic and intrinsic barriers is not enough . The third level barrier is the lack of design thinking, that is, instructors' ability to dynamically create knowledge and expertise in response to the pedagogical affordances of ICT (Tsai & Chai, 2012).

Realizing the potential of technology in education there are several ICT projects launched in India, to help and motivate learners. These include DIKSHA, Prerna, e-GyanKosh, NPTEL, e-PG

Pathshala, The National Literacy Mission, Sarva Shiksha Abhiyan (SSA) (Bist, 2007). However, the benefits of ICTs have not reached the expected level and maintaining the quality of education is one of the key challenges in India. Despite the resources like DIKSHA being available to the teachers, the implementation of the same is yet to be explored. An important gap is in the contextualization of these materials. Ultimately teachers have to teach with the same common set of materials in their own classes, for their target learners. Teachers' need to be able to take the commonly created materials and adapt them. Which means they need to build the necessary knowledge and skills to do this. Not sufficient attention has been given to capacity building of teachers or in improving teachers' learning design skills. In many developing countries, teacher professional development (TPD) programs need to be scaled so that ICT-based curricular materials can be effectively utilized. However, in teacher training programs different components are dealt with separately, due to which the technology, pedagogy and content area integration is a rare feature. Due to this, teachers were found to be weak in aligning the instructional activities and assessment for intended learning outcomes (i.e., they face some challenges in PCK). Thus, it is important to develop the TPACK knowledge of teachers to make the teaching and learning experience more meaningful. (Finger et al, 2015).

In this paper, our goal is to describe a workshop design *Learn to design* (L2D) and helping teachers implement research-based strategies using ICT based content, adapt it to their local context and enhance ease of use. The workshop design includes hands-on learning activities that focuses on enhancing the self-efficacy of teachers to integrate technology and practical adaptation of research based strategies into their teaching. Teachers can feel efficacious and in control of learning to teach with technology when their technology competency is improved. However, technology alone is not sufficient, pedagogical as well content knowledge is also crucial. The workshop activities are designed based on the TPACK framework (Mishra & Koehler, 2006), LCM model (Murthy et al, 2018) and constructive alignment (Biggs, 1996) to integrate technology, pedagogy and content. The two questions that are the focus of this study are as follows:

RQ 1. What are the effects of workshop training on participants' performance and perceived ability to design learning materials?

RQ 2. What are the effects of workshop training on participants' perception and their intention to use existing repositories of ICT materials?

2. Background and Related Work

Use of ICT helps in development of higher order thinking skills such as conceptual and procedural understanding, collaborating across time and place and real world problem-solving (Riess & Mischo, 2010). But the teachers are often unaware of ICT-enabled teaching methodology and effective integration strategies. A common issue is faced due to the teachers attitude and behaviour towards the use of ICT. Teachers' beliefs and attitudes towards the potentials of ICT in teaching and learning have been regarded as central conditions for successful implementation of new technologies. Even if a teacher has adequate resources, extensive digital instructional tools, and positive attitudes or strong ideas about technology integration, she/he may not be able to execute it successfully. (Dexter and Anderson, 2002, Ertmer et al., 1999, Newhouse, 2001). Another barrier that needs to be addressed for technology integration is the teachers' design thinking. Because the classroom environment and students are constantly changing, the teacher should use design thinking to reorganise or build learning materials and activities that adapt to the demands of diverse situations or groups of learners (Tsai and Chai, 2012).

Among the several ICT initiatives taken by the Govt of India DIKSHA Portal comprises curriculum aligned e-Learning content such as video lessons, assessments, worksheets and textbooks for all level students, teachers and even parents. It is a customizable platform currently being used by teachers (from both government and private institutions) across the nation. The platform offers various teacher training courses like Introduction to ICT, Integration of ICT in Teaching, Learning and Assessment and ICT tools. However, the synchronous interaction component is missing in these courses. This can be brought up by online workshops where teachers can interact with the instructor, perform simultaneous hands-on activities and receive real time feedback on the same. This helps in boosting the confidence and improving the self-efficacy of teachers that can be monitored. In order to facilitate this, we propose a workshop design (L2D) that can fulfil this gap.

3. Workshop Design

In order to provide training opportunities to teachers for effective implementation of ICT we have designed a workshop called *Learn to design(L2D).* The workshop design includes hands-on learning activities that focuses on enhancing the self-efficacy of teachers to integrate technology into their teaching. Teachers can feel efficacious and in control of learning to teach with technology when their technology competency is improved. However, technology alone is not sufficient and pedagogical as well as content knowledge is also crucial. The workshop activities are designed in such a way that they integrate technology, pedagogy and content. Table 1 describes the learning activities conducted in the workshop and their desired learning outcomes. In the activities for Day 1 of the workshop teachers were asked to choose a topic of their own choice for writing learning outcomes and respective assessment questions of different levels. For Day 2 teachers chose a particular video suitable for the previously chosen topic and edited it by adding reflection spot questions using a video editing software to make it suitable for their own context.

Table 1. *Session Plan and Outcomes of the Workshop*

Day 1- Synchronous mode			
Session title	Outcome (participants will be able to…..)	Activities	TPACK
1.1 Learning outcome why? What? & How?	Explain why learning objective is needed Identify valid learning objective	Writing learning outcomes	PK
1.2 Introduction to constructive alignment 1.3 Hands-on guided activity to write learning outcomes of different bloom's levels	Explain what is constructive alignment and why constructive alignment is needed. Classify the assessment questions as per Bloom's taxonomy. Generate questions for different levels of Bloom’s taxonomy. Align the assessment questions to their respective learning objectives.	Writing Assessment questions of Recall and understand level Apply and analyze level Evaluate and create level	PCK
Day 2 - Synchronous mode			
2.1 Introduction to LeD 2.2 Importance of Reflection spot 2.3 Hands on session on reflection spot insertion through technology tools (H5P, Vizia)	Explain components of LeD. Identify pause points in learning dialogue. Use of technology tools to chunk content and insert reflection spots. Create own learner centric learning dialogue in their context	Writing reflection spot questions along with feedback and timestamp.	TPACK

4. Implementation

We conducted 2 pilot workshops prior to doing the main workshop. The purpose of these pilots was to understand what components of the workshop need modifications. Due to the COVID-19 pandemic the workshops were conducted as synchronous online sessions spread over 2 days. The platform used for the workshop was Google meet. Workshop 1 was conducted with 15 in-service teacher participants from Maharashtra. In workshop 1 it was observed that the modules related to reflection spot activities had lower scores compared to the other activities with 33.3% participants scoring in a range that needs further improvement. Taking cues from this experience, we worked on the refinement of this module. With more time spent on the demonstration of the tool followed by hands-on activity. With these changes implemented, we conducted a second workshop with 27 in-service teacher participants from Maharashtra. We witnessed a positive outcome with the majority of participants performing well (50% scored valuable while 44.4% scored exemplar on a 3-point scale). With these insights from pilot workshop 1 and 2, a pre-workshop registration form was floated among the teacher groups in different states across the country. Total 308 participants participated in the workshop out of which 262 participants have been considered for data analysis in view of completeness of data available. We used Cisco Webex platform for conducting the workshop as it supports a large number of participants as compared to other video conferencing platforms available. It also provides a dial in option for participants to circumvent the internet bandwidth issue. Live synchronous sessions were planned for the duration of 2-2.5 hours on each day.

5. Methodology

We consider the sample of 262 participants who responded to the end of session survey and participated in the writing learning outcomes and assessment questions of different cognitive levels. There were a total of 9 data sources used for data collection. Pre-registration survey, pre-test, learning outcomes test, recall and understand level test, apply and analyze level test, evaluate and create level test, survey questionnaire on self-perception of constructive alignment and their intention to apply learning, reflection spot and feedback test and survey questionnaire on their self-perception on technology. The purpose of the pre-registration survey was to understand participants' familiarity with ICT and DIKSHA. A post-workshop survey questionnaire was used to capture the participant perceptions about the usefulness of this workshop, their engagement with the workshop modules, their perception about the use of technology and DIKSHA portal, and the pedagogical design of the workshop. The end of training survey for evaluating participant perception has been created and validated by the research team while the technology competence survey has been adapted from a standard instrument (Milman, Kortecamp & Peters, 2007). Participants' performance in the writing learning outcomes and assessment questions were evaluated on the basis of its specificity, measurability and use of appropriate action verbs. A 3-point scale rubric (1-potential to improve, 2-valuable, 3-exemplary) was used for evaluation. The reflection spots added by the participants were evaluated based on the chunking of content, suitability of the question and feedback provided for the learners.

6. Results

The survey results show that overall, there is a high perception of the training content, learning within the training and intention to apply among the participants. These changes in their perception are promising when compared to the pre-workshop data collected. The pre-workshop data shows that most of the participants (93.89%) were aware of the DIKSHA platform and 29% of them were regular users of DIKSHA. It was observed that 67.55% of them were restricted to the use of DIKSHA for training courses and 16.41% for content through either textbook QR code or mobile app. Only 9.16 % of the participants were making use of lesson plans provided on DIKSHA. Regarding the suitability of content available on DIKSHA, 19.46 % of the participants think that it requires certain modifications to make it suitable for their context. The pre-test results indicate that only 13.6% of the participants were able to identify valid learning outcomes and 35 % were able to identify assessment questions at different levels of Bloom's taxonomy from the given set of choices. The post workshop results show that participants were not only able to identify but also write valid learning outcomes and assessment questions. Participants' performance in the workshop learning activities on a 3-point scale suggest that 61.5% (161 out of 262) of the participants performed exemplary in writing learning outcomes while 76.7% (201 out of 262) were able to write exemplary assessment questions. Their performance in reflection spot activities suggest that 54.3 % (140 out of 262) of the participants have exemplary performance while 25.6% (66 out of 262) have valuable and 20.2 % potential to improve. Participants' perception of constructive alignment shows that ~95 % of them perceived to be confident about executing the constructive alignment between the learning outcomes and their corresponding assessment questions. The perception of participants on use of technology shows that ~40 % of the participants perceived themselves as capable of finding suitable video resources, using available resources for planning classroom activities, modifying the content as per the requirements and making use of tools for adding interactivity to the content on their own while ~31 % of them perceived that they can teach this to others. This change in perception also reflects in their intention to apply the learning from the workshop into practice. Intention to use the DIKSHA portal for lesson planning witnessed a significant rise with 91.2 % (Strongly Agree = 35.1%, Agree = 56.1%) participants intending to use DIKSHA.

7. Discussion

The effects of the workshop on participants' performance and perceived ability to design learning materials was evident from struggling to identify valid learning outcomes in the pre-test to writing exemplary learning outcomes in the learning activities. This implies that workshop activities contributed to their ability to design learning outcomes. This was also evident in the quality of assessment questions produced by the participants and their self-perceptions of the ability in terms of the constructive alignment. Participants designed assessments at different cognitive levels considering the diversity of their learners. There was clear alignment between the learning outcomes defined and the corresponding assessment questions designed for a particular topic.

For the effects of the workshop on participants' intention to use repositories of ICT materials we found that the workshop design helped in bringing a shift in attitude towards the use of DIKSHA, which is evident in their intention to use DIKSHA for activities such as lesson planning and structuring class activities. Participants perceived that they will be able to adapt and modify the available content to make it suitable for their context. Their ability to modify existing content was demonstrated in their artifacts produced in the form of interactive videos.

The study limitations include the short duration of the workshop (2-2.5 hours/day). We acknowledge that some of the activities may require more time to be spent for acquiring efficiency. For example participants struggled in writing assessment questions at higher levels of the bloom's taxonomy as compared to the lower levels. ~25% of the participants perceived that they need some assistance in using the technology. Spending some more time on the workshop learning activities may help to address these limitations. Due to time constraint the delayed post-test remains to be conducted and concluded. Thus we do not have any data about actual implementation of the learnings from the workshop in the classroom. We plan to conduct a delayed post-test with a subset of the participants to draw inferences about the impact of workshop learnings on classroom implementation.

The key contribution of our work is that we have provided a tested workshop design along with a detailed session plan and set of activities designed and curated for providing hands-on experience in online mode. This might be useful for teacher-trainers and educators, researchers working on teacher professional development . The workshop design can be replicated with some modifications as per the needs and context. The activities and tools have been explored to ensure a low entry barrier for the participants.

Acknowledgements

We would like to thank all the workshop participants and the organizing team.

References

Amhag, L., Hellström, L., & Stigmar, M. (2019). Teacher Educators' Use of Digital Tools and Needs for Digital Competence in Higher Education. Journal of Digital Learning in Teacher Education, 35(4), 203–220. https://doi.org/10.1080/21532974.2019.1646169

Biggs, J. (1996). Enhancing teaching through constructive alignment. Higher education, 32(3), 347-364. Bist, R.S. (2007), "ICT enabled development and digital divide: an Indian perspective", 5th International CALIBER 2007, Punjab University, Chandigarh, February 8-10, available at: http://ir.inflibnet.ac.in/bitstream/1944/1445/1/702-712.pdf (accessed March 16, 2017).

Brown, D. & Warschauer, M. (2006). From the University to the Elementary Classroom: Students' Experiences in Learning to Integrate Technology in Instruction. Journal of Technology and Teacher Education, 14(3), 599-621. Waynesville, NC USA: Society for Information Technology & Teacher Education. Retrieved May 30, 2021 from https://www.learntechlib.org/primary/p/5996/.

Dexter, S. L., Anderson, R. E., & Ronnkvist, A. M. (2002). Quality technology support: What is it? Who has it? And what difference does it make? Journal of Educational Computing Research, 26(3), 265–285

Ertmer, P. A. (1999). Addressing first-and second-order barriers to change: Strategies for technology integration. Educational Technology Research and Development, 47(4), 47–61.

Ertmer, P. A. (2005). Teacher pedagogical beliefs: The final frontier in our quest for technology integration?. *Educational technology research and development*, *53*(4), 25-39.

Finger, G., Jamieson-Proctor, R., Cavanagh, R., Albion, P., Grimbeek, P., Bond, T., ... Lloyd, M. (2013). Teaching Teachers for the Future (TTF) project TPACK survey: Summary of the key findings. Australian Educational Computing, 37(3), 13–25. https://espace.curtin.edu.au/handle/20.500.11937/2712

Gudmundsdottir, G. B., & Hatlevik, O. E. (2017). Newly qualified teachers' professional digital competence: implications for teacher education. European Journal of Teacher Education, 41(2), 214–231. https://doi.org/10.1080/02619768.2017.1416085

Howland, J. L., Jonassen, D. H., & Marra, R. M. (2012). Meaningful Learning with Technology. Pearson. Joo, Y. J., Lim, K. Y., & Kim, N. H. (2016). The effects of secondary teachers' technostress on the intention to use technology in South Korea. Computers & Education, 95(April 2016), 114–122. https://doi.org/10.1016/j.compedu.2015.12.004.

Koehler, M. J., & Mishra, P. (2005). What Happens When Teachers Design Educational Technology? The Development of Technological Pedagogical Content Knowledge. Journal of Educational Computing Research, 32(2), 131–152. https://doi.org/10.2190/0ew7-01wb-bkhl-qdyv

Lund, A., & Eriksen, T. M. (2016). Teacher Education as Transformation: Some Lessons Learned from a Center for Excellence in Education. Acta Didactica Norge, 10(2), 53–72. https://doi.org/10.5617/adno.2483 Mason, R. (2000). From distance education to online education. Internet and Higher Education, 3(1–2), 63–74. https://doi.org/10.1016/S1096-7516(00)00033-6

Milman, N. B., Kortecamp, K., & Peters, M. (2007). Assessing teacher candidates' perceptions and attributions of their technology competencies. International Journal of Technology in Teaching and Learning, 3(3), 15-35.

Murthy, S., Warriem, J., Sahasrabudhe, S., & Iyer, S. (2018). LCM: A model for planning, designing and conducting learner-centric MOOCs. Proceedings - IEEE 9th International Conference on Technology for Education, T4E 2018, 73–76. https://doi.org/10.1109/T4E.2018.00022

Newhouse, C. P. (2001). Applying the concerns-based adoption model to research on computers in class- rooms. Journal of Research on Computing in Education, 33(5), 2001.

Riess, W., & Mischo, C. (2010). Promoting systems thinking through biology lessons. International Journal of Science Education, 32(6), 705–725. https://doi.org/10.1080/09500690902769946.

Tsai, C. C., & Chai, C. S. (2012). The "third"-order barrier for technology-integration instruction: Implications for teacher education. Australasian Journal of Educational Technology, 28(6). https://doi.org/10.14742/ajet.810

Research on the Construction of Evaluation Indicators System of Pre-Service Teachers' Teaching Competency in Special Delivery Classroom

Xiangchun HE[a*], Peiliang MA[a,b*] & Xing ZHANG[a]
[a]*School of Educational Technology, Northwest Normal University, China*
[b]*Student Affairs Department, Tianshui Normal University, China*
*hxc@nwnu.edu.cn & mpl5697@163.com

Abstract: "Internet plus" pre-service teachers teaching-assistance is a new model of Internet-based education service proposed under the background of education informatization 2.0. At present, the *"Internet plus" Teaching-assistance Service Project for Normal College* has been implemented for two years as the breakthrough project for the director of *Gansu Provincial Department of Education*. However, with the deepening of the project, how to effectively evaluate the pre-service teachers' teaching competency of special delivery classroom has become a key issue. Based on this, the research focuses on providing scientific, effective and applicable tools and methods for the evaluation of teaching competency of pre-service teachers' in special delivery classroom. By comprehensively using the Literature Research Method, Delphi Method, Analytic Hierarchy Process (AHP) and Investigation Research Method, the evaluation indicators system of pre-service teachers' teaching competency in special delivery classroom with weight is constructed, It includes 5 first-level indicators, 12 second-level indicators and 36 third-level indicators. Which solves the key question of "what evaluation should be used?" The results show that the theoretically constructed evaluation indicators system has relatively comprehensive indicators, relatively reasonable weights, and strong applicability, and can objectively, comprehensively, fairly and justly evaluate the teaching competency of pre-service teachers in special delivery classroom.

Keywords: Special delivery classroom, pre-service teachers' teaching competency, evaluation indicators system

1. Introduction

Special delivery classroom has become one of the normal application ways to solve the teaching problems in rural small-scale schools, but it also puts forward new challenges for teachers. Combing and analyzing the projects that have been practiced at present, it is found that the teaching effect and quality are poor and can not be carried out sustainably for a long time due to the heavy task and low participation of teachers. "Internet plus" pre-service teachers teaching-assistance is mainly taught by pre-service teachers in normal schools after educational practice, and specialized classes for rural small-scale school students through special delivery classroom. It not only effectively cracked the difficulties in offering courses in the state, but also broadened the channels for pre-service teachers' education practice, and promoted their teaching competency to progressively improve (Guo, 2020). Put forward the "Gansu Program" to promote the balanced development of high quality education.

Therefore, according to the practical demands, trying to construct an evaluation indicators system that can be used to evaluate the pre service teachers' teaching competency of special delivery classroom is very necessary and has important practical significance and value. Based on the above analysis, the core issues of research is put forward: how to construct a scientific and reasonable evaluation indicators system? Around the core issues, we need to solve the following specific issues:

1. How to determine the evaluation dimensions and indicators?
2. How to determine the importance of indicators?

2. Related Literature and Theoretical Foundations

2.1 Overview of Special Delivery Classroom

Professor Wang J's team started to carry out the "Internet plus" teaching site Hubei action in 2014, proposed "Internet plus" localized classrooms (Wang, J. 2016). Built a central school with M teaching sites (M is generally takes the value between 1 to 3) N (1+M) model of N localized teaching communities, which "synchronized interactive delivery classroom" is one of the specific teaching models (Tian, 2019). Lei L and Zuo M constructed structural framework of synchronous interactive hybrid classroom teaching model for rural teaching sites (Lei, 2015). Shen J, Guo S and other scholars designed a special delivery classroom teaching model based on a broadband satellite network environment and carried out practical applications (Guo, 2020). In addition, a large number of domestic education and training institutions and non-profit organizations have applied special delivery classroom teaching model extensively.

In addition, foreign studies similar to special delivery classroom mainly include: based on the analysis of the history of distance education in primary and secondary schools in North America, Michael Barbour proposed methods and ways to explore the construction of rural virtual schools from the perspective of online learning, hoping to improve the quality of rural school education. (Barbour, 2011). Alabama has built an online and interactive video system linking teachers and students across the state. Summerdale School uses online synchronous or asynchronous classes to realize that teachers can remotely teach students from remote rural teaching sites in the central school, which is a typical representative (Wang, X. 2016).

In summary, although there are many research results in the application model of special delivery classroom, they mainly focus on the teaching model, interactive behavior, problem analysis, optimization strategy, influencing factors and effect evaluation. However, the research on the teaching competency under the technical environment of special delivery classroom is relatively weak.

2.2 Overview of Pre-Service Teachers' Teaching Competency Evaluation Indicators System

In order to diagnose the level of pre-service teachers teaching competency, a diagnostic tool corresponding to the *"Pre-service Teacher Informatization Teaching Competency Standards"* is developed, which is proved to has good reliability and validity through multiple rounds of iterations. It is the main tool for the evaluation of pre-service teachers' informatization teaching competency in China (Yan, 2015).

In recent years, researchers have begun to pay attention to the evaluation of pre-service teachers' teaching competency. Different scholars have discussed the evaluation indicator system of pre-service teachers' teaching competency from different perspectives. By analyzing the teaching process of pre-service teachers, Han G has constructed the evaluation indicators system of pre-service teachers teaching competency through expert evaluation and self-evaluation (Han, 2011). Wang H used the literature research method to construct an evaluation system of physics pre-service teachers' teaching skills and applied the evaluation system in practice by issuing questionnaires (Wang, H. 2015). Pu C constructed the evaluation indicators system of pre-service teacher teaching competency, obtained the optimal weight coefficient based on game theory to combine weighting, and carry out empirical research to verify its effectiveness (Pu, 2019). In addition, edTPA is a typical pre-service teacher's teaching competency evaluation system, which comprehensively evaluates the teaching plans, teaching reflections, teaching videos, so as to reflect the level of competency (American Association of Colleges for Teacher Education, 2014).

To sum up, most of the existing evaluation indicators system are formulated according to relevant standards, but there are still some shortcomings. For example, some studies are relatively simple in the selection and extraction of indicators and the calculation methods of weights, and some studies only construct the evaluation indicators system at the theoretical level, but do not verify its rationality and effectiveness. The direct selection of samples for actual evaluation application cannot guarantee the reliability and validity of the research. In addition, the research on the teaching competency evaluation indicators system mainly focuses on in-service teachers, and the research results on pre-service teachers are relatively weak.

3. Methods and Procedures

3.1 Literature Research Method

This study mainly uses the literature research method to understand and master the relevant research results and materials on the special delivery classroom, and pre-service teachers' teaching competency evaluation indicators system at home and abroad. so as to find the entry point and reference basis for establishing evaluation indicators at all levels and preliminarily drafting the pre-service teachers' teaching competency evaluation indicators system.

3.2 Delphi Method

Delphi Method mainly solicits the opinions of experts in the field by preparing a questionnaire. After repeated letters and feedback, the evaluation index system is iteratively revised. Finally, the opinions of experts gradually converge and obtain consistent opinions.

3.3 Analytic Hierarchy Process

AHP is a hierarchical decision analysis method that decomposes the overall goal into various indicators at different levels, and then constructs pairwise comparison judgment matrix at each level to test the consistency, compare the relative importance and determine the weight value. After determining the evaluation indicators system through multiple rounds of expert consultation, yaahp is used to establish a hierarchical structure model, and the weight determination expert questionnaire is distributed in the "Yue ping" comprehensive evaluation service platform. After collecting the expert decision-making data, it is imported into yaahp, automatically constructs the judgment matrix and carries out consistency test, and then calculates the weight value of each indicators.

3.4 Construction Process and Perspective

Pre-service teachers' teaching competency in special delivery classroom are a kind of comprehensive teaching competency presented as a whole. To construct an evaluation indicators system to evaluate this comprehensive teaching competency, it is necessary to analyze and deconstruct it from an all-round and multi angle, and screen out evaluation indicators from different perspectives. According to the theory of teaching competency structure in the previous research, and referring to the micro certification project of teachers' information technology application competency in East China Normal University, this study decomposes the competency from the perspective of organization, teaching and competency, and comprehensively considers the requirements of teachers' role, teaching situation and the definition of teachers' competency. The construction of evaluation indicators system is not a simple combination of evaluation indicators, but a systematic work with purpose and hierarchy. It is also a process of iterative optimization, which is usually completed through several steps as shown in Figure 1.

4. Results

Based on literature review and analysis, following the principles and basis, ideas and methods of constructing evaluation indicators system, this study initially constructs evaluation indicators system of pre-service teachers' teaching competency in special delivery classroom from the three perspectives of pre-service teacher, special delivery classroom and teaching competency. Then, the expert consultation questionnaire was compiled. Following the principle of combining authority and representativeness, experts in the field of educational informatization and teacher education were invited for consultation, and three rounds of expert consultation questionnaires were issued for expert opinions. After the first round of expert consultation, the indicators at all levels have been modified and adjusted to varying degrees. After the second round of expert consultation, the opinions of the expert group basically tend to be consistent.

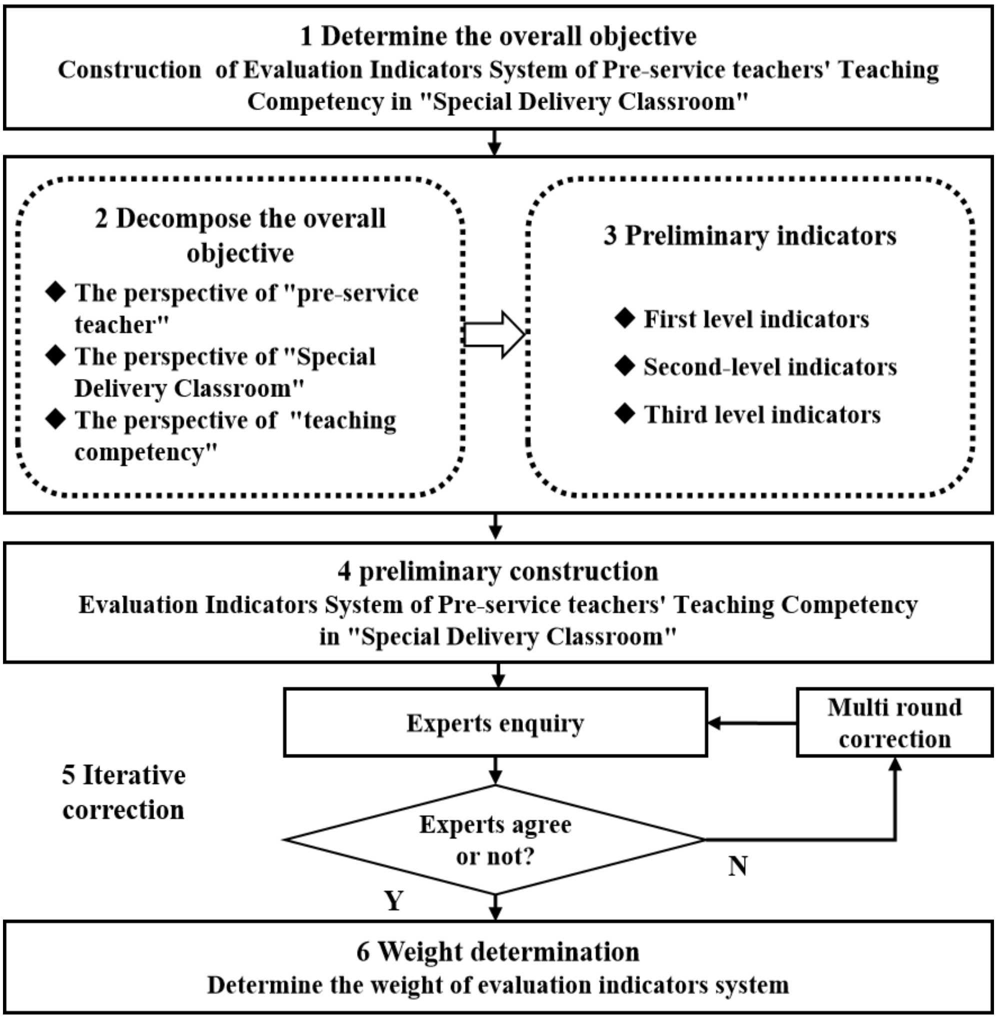

1 Determine the overall objective
Construction of Evaluation Indicators System of Pre-service teachers' Teaching Competency in "Special Delivery Classroom"
2 Decompose the overall objective
◆ The perspective of "pre-service teacher"
◆ The perspective of "Special Delivery Classroom"
◆ The perspective of "teaching competency"
3 Preliminary indicators
◆ First level indicators
◆ Second-level indicators
◆ Third level indicators
4 preliminary construction
Evaluation Indicators System of Pre-service teachers' Teaching Competency in "Special Delivery Classroom"
Experts enquiry
Multi round correction
5 Iterative correction
Experts agree or not?
N
Y
6 Weight determination
Determine the weight of evaluation indicators system

Figure 1. The Construction Process of Evaluation Indicators System.

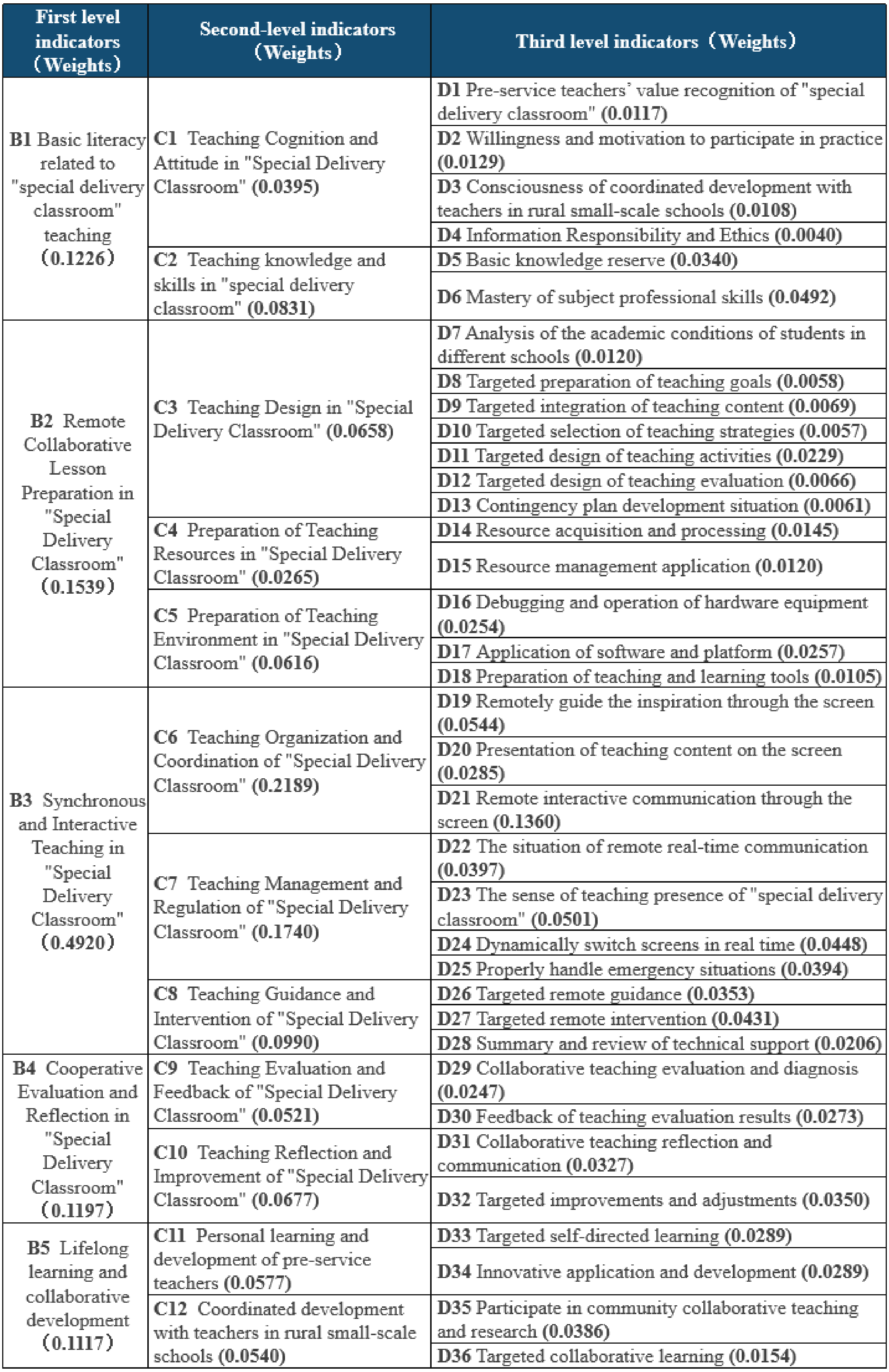
Weight is the relative importance of an indicators in the whole indicators system. In the third round, the weights of indicators are determined by the combination of Delphi Method and AHP. Firstly, the hierarchical structure model is constructed by yaahp, and then "Yue ping" comprehensive evaluation service platform is used to collect the expert decision-making data, which is imported into yaahp to automatically construct the judgment matrix, and the weight values of indicators at all levels are calculated by group decision-making. Finally, evaluation indicators system of pre-service teachers' teaching competency in special delivery classroom with weight is obtained as shown in Figure 2.

5. Discussion

Due to the limited space, the following is based on the scoring results of the expert group, and only takes the determination of the weight of the first level indicators as an example. In yaahp, based on the hierarchical model, and the experts' scoring results at each level are obtained. The details are shown in Table 1. At the same time, the weight ranking is obtained, as shown in Table 2.

Table 1. *Summary of Expert Consultation Results of First-Level Evaluation Indicators*

A	B1	B2	B3	B4	B5	Wi
B1	1	0.7965	0.2492	1.0238	1.0974	0.1226
B2	1.2555	1	0.3129	1.2854	1.3778	0.1539
B3	4.0132	3.1964	1	4.1087	4.4040	0.4920
B4	0.9768	0.7780	0.2434	1	1.0719	0.1197
B5	0.9113	0.7258	0.2271	0.9329	1	0.1117



First level indicators（Weights）	**Second-level indicators（Weights）**	**Third level indicators（Weights）**
B1 Basic literacy related to "special delivery classroom" teaching **（0.1226）**	**C1** Teaching Cognition and Attitude in "Special Delivery Classroom" **(0.0395)**	**D1** Pre-service teachers' value recognition of "special delivery classroom" **(0.0117)**
		D2 Willingness and motivation to participate in practice **(0.0129)**
		D3 Consciousness of coordinated development with teachers in rural small-scale schools **(0.0108)**
		D4 Information Responsibility and Ethics **(0.0040)**
	C2 Teaching knowledge and skills in "special delivery classroom" **(0.0831)**	**D5** Basic knowledge reserve **(0.0340)**
		D6 Mastery of subject professional skills **(0.0492)**
B2 Remote Collaborative Lesson Preparation in "Special Delivery Classroom" **（0.1539）**	**C3** Teaching Design in "Special Delivery Classroom" **(0.0658)**	**D7** Analysis of the academic conditions of students in different schools **(0.0120)**
		D8 Targeted preparation of teaching goals **(0.0058)**
		D9 Targeted integration of teaching content **(0.0069)**
		D10 Targeted selection of teaching strategies **(0.0057)**
		D11 Targeted design of teaching activities **(0.0229)**
		D12 Targeted design of teaching evaluation **(0.0066)**
		D13 Contingency plan development situation **(0.0061)**
	C4 Preparation of Teaching Resources in "Special Delivery Classroom" **(0.0265)**	**D14** Resource acquisition and processing **(0.0145)**
		D15 Resource management application **(0.0120)**
	C5 Preparation of Teaching Environment in "Special Delivery Classroom" **(0.0616)**	**D16** Debugging and operation of hardware equipment **(0.0254)**
		D17 Application of software and platform **(0.0257)**
		D18 Preparation of teaching and learning tools **(0.0105)**
B3 Synchronous and Interactive Teaching in "Special Delivery Classroom" **（0.4920）**	**C6** Teaching Organization and Coordination of "Special Delivery Classroom" **(0.2189)**	**D19** Remotely guide the inspiration through the screen **(0.0544)**
		D20 Presentation of teaching content on the screen **(0.0285)**
		D21 Remote interactive communication through the screen **(0.1360)**
	C7 Teaching Management and Regulation of "Special Delivery Classroom" **(0.1740)**	**D22** The situation of remote real-time communication **(0.0397)**
		D23 The sense of teaching presence of "special delivery classroom" **(0.0501)**
		D24 Dynamically switch screens in real time **(0.0448)**
		D25 Properly handle emergency situations **(0.0394)**
	C8 Teaching Guidance and Intervention of "Special Delivery Classroom" **(0.0990)**	**D26** Targeted remote guidance **(0.0353)**
		D27 Targeted remote intervention **(0.0431)**
		D28 Summary and review of technical support **(0.0206)**
B4 Cooperative Evaluation and Reflection in "Special Delivery Classroom" **（0.1197）**	**C9** Teaching Evaluation and Feedback of "Special Delivery Classroom" **(0.0521)**	**D29** Collaborative teaching evaluation and diagnosis **(0.0247)**
		D30 Feedback of teaching evaluation results **(0.0273)**
	C10 Teaching Reflection and Improvement of "Special Delivery Classroom" **(0.0677)**	**D31** Collaborative teaching reflection and communication **(0.0327)**
		D32 Targeted improvements and adjustments **(0.0350)**
B5 Lifelong learning and collaborative development **（0.1117）**	**C11** Personal learning and development of pre-service teachers **(0.0577)**	**D33** Targeted self-directed learning **(0.0289)**
		D34 Innovative application and development **(0.0289)**
	C12 Coordinated development with teachers in rural small-scale schools **(0.0540)**	**D35** Participate in community collaborative teaching and research **(0.0386)**
		D36 Targeted collaborative learning **(0.0154)**

Figure 2. Evaluation Indicators System (with weight).

Table 2. *First-Level Evaluation Indicators Weight Ranking*

First-level indicators	Weight	Weight ranking
B3	0.4920	1
B2	0.1539	2
B1	0.1226	3
B4	0.1197	4
B5	0.1117	5

6. Conclusion

This study mainly constructs the evaluation indicators system of pre-service teachers' teaching competency in special delivery classroom with weight, and verifies its scientificity, rationality and operability, so as to solve the key problem of "what evaluation should be used?" The results show that the evaluation indicators system constructed by the theoretical method is relatively comprehensive, the weight is relatively reasonable and has strong applicability, which can objectively, comprehensively, fairly and justly evaluate the pre-service teachers' teaching competency in special delivery classroom.

In a word, the research has made some achievements in exploring the theory and method of pre-service teachers' teaching competency evaluation in special delivery classroom, but it still needs to be optimized and improved in the future practice, and deepen and expand the evaluation indicators system for different subjects, different schools and different disciplines. So that the pre-service teachers teaching competency in special delivery classroom evaluation is more standardized, systematic, and gradually toward accuracy and personalization.

Acknowledgements

Project Research on Key Technologies of data literacy intelligent evaluation of primary and secondary school teachers based on multi-source information fusion supported by NSFC (No.62167007).

References

American Association of Colleges for Teacher Education. 2014 edTPA administrative report [EB/OL]. [2016-12 -25]. *https://secure.aacte.org/apps/rl/resource.php?resid=558&ref=edtp*.

Barbour, M. (2011). The promise and the reality: exploring virtual schooling in rural juristictions. *Education in Rural Australia. 21*(01):1-19.

Guo, S., Shen, J., Zhang, J., & He, X. (2020). Research on the "Five in One" Collaborative Teaching Research Model under the Condition of "Internet+". *e-Education Research. 41*(12):35-42.

Han, G. (2011). Research on the evaluation of free normal students' teaching ability. *Shaanxi Normal University*.

Lei, L., Zuo, M. (2015). A Hybrid Learning Model for Synchronous Interaction Classroom Oriented towards Rural Schools. *e-Education Research*. *36*(11):38-43.

Pu, C. (2019). The evaluation of normal students' informatization teaching ability based on the combination weighting of game theory. *Chongqing Normal University*.

Tian, J., Wang, J., & Wang, X. (2019). "internet + localization": a practical study on the improvement of teaching quality in rural schools. *China Educational Technology*. (10):38-46.

Wang, J., Feng, S., & Wu, X. (2016). Internet plus rural small-schools: practice research on the balanced development of compulsory education in the process of new urbanization. *China Educational Technology*. (01):86-94.

Wang, X., Zhang, J., Jing, D., Zhang, D., & Gao, Y. (2016). Predicting the development of educational technology and exploring related instructional model under global ICT context —— a review of us-china smart education conference. *e-Education Research. 37*(03):34-41.

Wang, H. (2015). Research on the construction and application of teaching skill evaluation system for physics normal students. *Contemporary Educational Science.* (13): 20-23.

Yan, H., Li, X., & Ren, Y. (2018). Development and validation of self-measurement tools for pre-service teachers' ICT competency. *e-Education Research. 39*(01):98-106.

Students with Disabilities and Digital Accessibility in Higher Education under COVID-19

Weiqin CHEN[a,b*]
[a]*Department of Computer Science, Oslo Metropolitan University, Norway*
[b]*SLATE, University of Bergen, Norway*
*weiche@oslomet.no

Abstract: The COVID-19 pandemic has disrupted education system worldwide. The emergency remote teaching has affected higher education institutions and created challenges for multiple stakeholders including students, faculties, administrators, and policy makers. In this context, the challenges faced by students with disabilities (SWDs) worsened by COVID-19. This paper considers and reflects on existing research on experiences of students with disabilities under the COVID-19 pandemic and identifies challenges and opportunities for supporting students with disabilities and implementing digital accessibility in higher education in post-pandemic world.

Keywords: Students with disabilities, digital accessibility, COVID-19, higher education

1. Introduction

With the increased prevalence of the right to equal access to education in many countries, more and more students with disabilities (SWDs) are taking higher education. According to the 2018 European Student Survey (Hauschildt, Vögtle, & Gwosć, 2018), an average 18% of students in higher education reports having a disability or chronic disease. In the US, the number of students with disabilities in colleges was estimated to be 19% in 2015/2016 (Snyder, de Brey, & Dillow, 2019).

The COVID-19 pandemic has disrupted education system worldwide affecting over 94% of the world’s student population (UN, 2020). The emergency remote teaching has created challenges for all stakeholders in higher education, particularly for students with disabilities. The technological barriers with remote teaching such as inaccessible teaching materials and platforms and the socio-economic challenges such as lack of access to necessary resources have negatively affected the learning experiences of students with disabilities. For example, lectures videos without captioning prevents deaf and hard-of-hearing students from access without sign language translation. Blind and visually impaired students reply on text or audio descriptions of informative images in lecture notes.

Previous research has shown that SWDs were at greater risk of prematurely withdrawing from university or dropping out compared to students without disabilities (Lombardi, Murray, & Kowitt, 2016). Although some of the challenges have already been identified pre-COVID and measures have been taken to facilitate the success of SWDs in many higher education institutions, the emergency remote teaching under COVID-19 has posed greater challenges on faculties and administration in higher education. This has also resulted in that the needs of SWDs were not prioritized or neglected and resources needed by SWDs were not sufficiently allocated. SWDs has therefore become a progressively vulnerable group whose educational needs were not addressed under the COVID-19 pandemic.

In this paper, we review existing research on experiences of SWDs and digital accessibility under COVID-19. We argue that the COVID-19 pandemic does not only create challenges for SWDs in higher education, but the disruption also stimulates research and innovation in supporting SWDs and implementing digital accessibility, which has implications for the post-pandemic world.

2. Previous Research on SWDs in Higher Education

A large number of research pre-COVID have focused on challenges of SWDs face in higher education (Fuller, Healey, Bradley, & Hall, 2004; Healey, Bradley, Fuller, & Hall, 2006; Madriaga et al., 2010) and revealed that SWDs invest more time and effort than their non-disabled peers do in coping with challenges (Berggren, Rowan, Bergbäck, & Blomberg, 2016; Goode, 2007).

Marquis and colleagues (Marquis, Fudge Schormans, et al., 2016; Marquis et al., 2012; Marquis, Jung, et al., 2016) present a three-phase qualitative study focusing on accessible education for SWDs in a Canadian University. The study interviewed students with and without disabilities, instructors, staff members and administrators and identified factors affecting educational accessibility in the university. These factors include knowledge, attitude, pedagogical choices, disciplinary features and institutional practices and characteristics. For example, instructors' lack of knowledge about disabilities, teaching and learning strategies employed by instructors such as group work, certain kinds of discussion, and lecturing without visual/textual support, and inaccessible course materials were reported as barriers by students with disabilities. More recently, Jeannis et al. (2019) conducted a national self-report survey on students with physical disabilities in science and engineering education and identified barriers including inappropriate accommodations and instructors' negative viewpoints. The empirical data also revealed a range of facilitators including accessible course materials and peer assistance.

There is also a large body of research on Universal Design for Learning (UDL), a framework with a set of principles for curriculum development that address barriers and give all students equal opportunities to learn. The key concepts underlying UDL are adopting multiple means of content delivery, diverse methods of expression and assessment, and different means of engagement (Rose & Meyer, 2002).

3. SWDs under the COVID-19 Pandemic

Under COVID-19, several researchers have conducted studies focusing on experiences of students with disabilities and digital accessibility in higher education. Through searching in Google Scholar with general keywords such as disability, COVID-19, higher education, and more specific keywords such as blind, visual impairment, deaf, hard-of-hearing, hearing impairment, learning disability, intellectual disability, and motor disability, we have identified and collected the relevant publications. The search was conducted between 20th and 26th April 2021. In this section, we present the analysis of the collected publications.

3.1 Studies Focusing on Challenges of SWDs under COVID-19

Except for Krishnan et al. (2020) which focused on students with hearing impairments, all the other studies in Table 1 covered challenges of students with diverse disabilities, both visible such as physical impairments and invisible such as mental disorder and learning disabilities. These studies have identified a number of challenges SWDs face under COVID-19.

Table 1. *Overview of the Studies on SWDs Challenges*

Study	Focused disability	Study method	Participants	Country
(Bartz, 2020)	General, mental disorders	Survey, interview	Students (n=45)	Germany
(Krishnan et al., 2020)	Hearing impairment	Interview	Student (n=10)	Malaysia

(Pichette, Brumwell, & Rizk, 2020)	General	Survey, interview	Students (survey n=623, 208 with a disability; interview 11 students)	Canada
(Scott & Aquino, 2021)	General	Survey	disability Professionals (n= 645)	USA
(Soria, Horgos, Chirikov, & Jones-White, 2020)	General	Survey	Students (n = 30 099, 1 788 with a disability)	USA
(Zhang et al., 2021)	General	Survey	Students (n=147, 28 with disabilities/health concerns)	USA

Apart from Scott and Aquino (2021) which collected challenges and barriers based on experiences of disability professionals who are responsible for accommodating SWDs in universities, all the other studies collected data directly from students with disabilities. The challenges and barriers can be grouped into three main categories:

- ***Infrastructure*** including lack of access to network/WiFi and assistive technology, lack of digital literacy, lack of access to equipment for online learning. For example, deaf and hard-of-hearing students experienced a lack of sign language interpreters. Students with visual impairments experienced a lack of competence in using screen readers.
- ***Learning material and platform*** including difficulty accessing learning material and using learning platforms, lack of access to course exams or assessment, and difficulty accessing technical support. For example, deaf and hard-of-hearing students experienced a lack of captioning for videos and live events. Students with visual impairments experienced a lack of descriptions for informative images in teaching notes, and difficulties in using learning management systems.
- ***Communication*** including difficulty communicating with support services, faculty/instructors and fellow students due to for example lack of privacy when SWDs live in family homes (Pichette et al., 2020).

In addition, Pichette et al. (2020), Soria et al. (2020) and Zhang et al. (2020) all reported SWDs having difficulties focusing and concentrating, experiencing distress, and other psychological challenges during COVID-19. In addition, Soria et al. (2020) also reported financial challenges, food and housing insecurity and safety challenges. Focusing on hearing impairments, Krishnan et al. (2020) identified challenges including accessing to hearing devices, difficulty in following and understanding lessons, and unfamiliarity with online services, as well as psychological distress.

3.2 Studies Focusing on Challenges of Faculty and Good Teaching Practice

Challenges in providing accessibility for SWDs faced by faculty under COVID-19 were highlighted by studies in Table 2. These challenges include understanding and satisfying the needs of SWDs, providing accessible curricula, pedagogical design, teaching materials, and assessment, lack of adequate support, guidance, and training for working with SWDs in online teaching. In some subjects such as surgery (Dickinson & Gronseth, 2020), human anatomy (Guimarães, Lima, Teixeira, Sanchez, & Mendonça, 2020; Mendonça, Souza, Arruda, Noll, & Guimarães, 2021; Pacheco, Noll, & Mendonça, 2020), chemistry (Lynn et al., 2020), and program programming (Li et al., 2021), the shift to online teaching posed challenges for faculty to accommodate students with hearing, visual impairments, and learning and intellectual disabilities, particularly when faculties themselves were also under stress in adapting to online teaching. The studies (Dickinson & Gronseth, 2020; Guimarães et al., 2020; Li et al., 2021; Lynn et al., 2020; Mendonça et al., 2021; Pacheco et al., 2020) present good practices in teaching the specific subjects to students with specific disabilities. Behling (2020) and Dickinson & Gronseth (2020) suggested adopting Universal Design for Learning (UDL) principles to provide inclusive education under COVID-19 to ensure access to online learning for all students. In addition to surveying and interviewing students, Pichette et al. (2020) in their study also interviewed four instructors, identified their challenges, and proposed recommendations for instructors, one of which is incorporating UDL principles in all courses.

Table 2. *Overview of the Studies on Faculty Challenges and Good Teaching Practice*

Study	Focused disability	Subject	Country
(Behling, 2020)	General	General	USA
(Dickinson & Gronseth, 2020)	General	Surgery	USA
(Guimarães et al., 2020)	Hearing impairments	Human anatomy	Brazil
(Li et al., 2021)	Learning Disabilities	Pair programming	USA
(Lynn et al., 2020)	Hearing impairments	Chemistry	USA
(Mendonça et al., 2021)	Low vision and blind	Human anatomy	Brazil
(Pacheco et al., 2020)	Intellectual disability	Human Anatomy	Brazil

3.3 Studies Focusing on Challenges of Support Services and Good Practice

Table 3 shows an overview of the studies on challenges of support services and good practice. The studies by Pichette et al. (2020) and Scott & Aquino (2021) both highlighted the challenges faced by disability service staff in accommodating SWDs in remote learning environments under COVID-19. Scott and Aquino (2021) reported that a large number of students in need of accommodation, lack of access to equipment/devices, difficulties in communicating with students, faculty and other departments and offices, lack of institutional support and funding have affected their services to SWDs. The study further identified institutional strategies for supporting SWDs under COVID-19, including involvement in policy development, participation in campus structure and committees, provision of training and instructions, and collaboration with other departments.

Table 3. *Overview of the Studies on Challenges of Support Services and Good Practice*

Study	Study method	Participant	Country
(Lazar, 2020)	Interview	Directors of digital accessibility (n=18)	USA
(Meleo-Erwin et al., 2021)	Web content analysis	University websites (n=127)	USA
(Pichette et al., 2020)	Survey, interview	Disability services staff (survey n=72, 208; interview 12 staff)	Canada
(Scott & Aquino, 2021)	Survey	Disability Professionals (n= 645)	USA

Meleo-Erwin et al. (2021) focused on disability/accessibility pages on university websites and investigated whether these pages provide links to instructional resources for students, faculty members, counselling center and possibility to make an appointment with the counselling center. The results indicated that there was a lack of key information on copying with the challenges wrought by COVID-19 on the disability/accessibility page. The study further suggested that visible and accessible links to remote instruction resources and counselling services be places in the disability/accessibility pages of all higher education institutions.

Lazar (2020) conducted interviews with directors of digital accessibility in US universities to understand their strategies and challenges concerning digital accessibility under the COVID-19 pandemic. The findings showed that universities bypassed accessibility requirements in their emergency procurement and purchased inaccessible digital technologies under swift shift to remote teaching. Some of the inaccessible technologies will be used even after COVID-19 and therefore will have long-term consequence for SWDs. Furthermore, captioning of videos and live events was found to be a challenge due to lack of funding. Some universities use automatic captioning for example in Zoom despite of low quality and only provide live captioning upon request. The lack of resources and qualified personals created barriers for students with hearing impairments. PDF accessibility was also found to be an issue since many faculty and administrations were reported to scan paper documents and forms to image files and send them via email or upload them to an online platform to share with students. This could be caused by lack of awareness of the challenges faced by SWDs and/or lack of digital accessibility competence among faculty and administration. It could also be that faculty and administration do not have enough time to work out accessible solutions. Based on the analysis of the interviews, Lazar (2020) made recommendations to address identified challenges related to procurement, document accessibility, accessibility training, and captioning. Lazar (2020) further suggested that universities need to make plans to prepare for “surge” in demand of accessibility-related services caused by emergency situations such as COVID-19.

4. Discussion and Conclusion

Students with disabilities and digital accessibility suffer during the COVID-19 pandemic. This has been evidenced by several published studies as well as newspaper articles. Although some of the challenges SWDs face are well-documented problems in pre-COVID time, the emergency remote teaching has exacerbated the problems, and caused/is causing more severe consequences for SWDs. Such consequences will not only affect SWDs during, but also long after the COVID-19 pandemic. The digital tools procured during COVID-19 bypassing accessibility requirements are one of the examples of such consequences (Lazar, 2020).

Despite of a growing body of research addressing the challenges of SWDs in higher education and proposing strategies for inclusive education, there is still a gap between what SWDs need and what universities including faculties, administrative staff and leadership do to meet the needs. This gap is further widened by the emergency transferring to digital teaching due to COVID-19.

It is positive to see that some of the universities represented in the literature have increased awareness and are continuously improving their practices in supporting SWDs and implementing digital accessibility under COVID-19. However, the limited literature related to SWDs under COVID-19 has also demonstrated that SWDs and digital accessibility have not received enough attention in higher education. Furthermore, most of the studies were conducted in high-income countries. Further research should focus on low- and middle-income countries. In addition, research should also investigate the experiences of SWDs and experiences of faculty and support services in accommodating SWDs in different disciplines. Such research may provide a better understanding of the challenges faced by SWDs in higher education and good practice in addressing the challenges.

Through literature analysis, this paper has shown that a clear policy at the university level which reflected in the procurement of digital technologies, supporting services and resources allocation is necessary for better accommodate SWDs. Furthermore, capacity building, including providing incentives and training of both faculty and administrative staff to increase awareness of SWDs and digital accessibility and know-how in creating accessible teaching is important to implement digital accessibility in universities, which will further contribute to an equitable learning experience for SWDs.

References

Bartz, J. (2020). All Inclusive?! Empirical Insights into Individual Experiences of Students with Disabilities and Mental Disorders at German Universities and Implications for Inclusive Higher Education. *Education Science, 10*. Retrieved from https://files.eric.ed.gov/fulltext/EJ1271961.pdf

Behling, K. (2020). Finding a silver lining in the rapid movement to online learning: Considerations of access for all learners. *Pedagogy and the Human Sciences, 7*(1), 1-11.

Berggren, U. J., Rowan, D., Bergbäck, E., & Blomberg, B. (2016). Disabled Students’ Experiences of Higher Education in Sweden, the Czech Republic, and the United States – A Comparative Institutional Analysis. *Disability & society, 31*(3), 339-356. doi:doi:10.1080/09687599.2016.1174103

Dickinson, K. J., & Gronseth, S. L. (2020). Application of Universal Design for Learning (UDL) Principles to Surgical Education During the COVID-19 Pandemic. *Journal of Surgical Education, 77*(5), 1008-1012. doi:https://doi.org/10.1016/j.jsurg.2020.06.005

Fuller, M., Healey, M., Bradley, A., & Hall, T. (2004). Barriers to learning: A systematic study of the experience of disabled students in one university *Studies in Higher Education, 29*, 303-318. doi:http://dx.doi.org/10.1080/03075070410001682592

Goode, J. (2007). ‘Managing’ Disability: Early Experiences of University Students with Disabilities. *Disability & society, 22*(1), 35-48.

Guimarães, N., Lima, B. S., Teixeira, A. C., Sanchez, B. Z., & Mendonça, C. R. (2020). Difficulties encountered by hearing impaired students, teachers and interpreters of the Brazilian sign language in teaching-learning human anatomy in higher education courses. *Reseach, Society and Development, 9*(6). doi:https://doi.org/10.33448/rsd-v9i6.3478

Hauschildt, K., Vögtle, E. M., & Gwosć, C. (2018). *Social and Economic Conditions of Student Life in Europe. EUROSTUDENT VI 2016–2018 | Synopsis of Indicators.* : German Centre for Higher Education Research and Science Studies (DZHW).

Healey, M., Bradley, A., Fuller, M., & Hall, T. (2006). Listening to students: The experiences of disabled students of learning at university. In M. Adams & S. Brown (Eds.), *Towards inclusive learning in higher education: Developing curricula for disabled students* (pp. 32-43). London: Routledge.

Jeannis, H., Goldberg, M., Seelman, K., Schmeler, M., & Cooper, R. A. (2019). Barriers and facilitators to students with physical disabilities' participation in academic laboratory spaces. *Disability and Rehabilitation: Assistive Technology, 15*(2), 225-237.

Krishnan, I., Mello, G., Kok, S., Sabapathy, S., Munian, S., Ching, H., . . . Kanan, V. (2020). Challenges Faced by Hearing Impairment Students During COVID-19. *Malaysian Journal of Social Sciences and Humanities (MJSSH), 5*(8), 106-116.

Lazar, J. (2020). Managing digital accessibility at universities during the COVID-19 pandemic. *Univ Access Inf Soc*. doi:https://doi.org/10.1007/s10209-021-00792-5

Li, L., Xu, L., He, Y., He, W., Pribesh, S., Watson, S. M., & Major, D. A. (2021). Facilitating Online Learning via Zoom Breakout Room Technology : A Case of Pair Programming Involving Students with Learning Disabilities. *Communications of the Association for Information Systems, 48*. Retrieved from https://aisel.aisnet.org/cais/vol48/iss1/12/

Lombardi, A., Murray, C., & Kowitt, J. (2016). Social Support and Academic Success for College Students with Disabilities: Do Relationship Types Matter? *Journal of Vocational Rehabilitation, 44*, 1-13.

Lynn, M. A., Templeton, D. C., Ross, A. D., Gehret, A. U., Bida, M., Sanger, T. J., & Pagano, T. (2020). Successes and challenges in teaching chemistry to deaf and hard-of-hearing students in the time of COVID-19. *J. Chem. Educ.* . doi:10.1021/acs.jchemed.0c00602

Madriaga, M., Hanson, K., Heaton, C., Kay, H., Newitt, S., & Walker, A. (2010). Confronting similar challenges? Disabled and non-disabled students' learning and assessment experiences. *Studies in Higher Education, 35*(6), 647-658. doi:http://dx.doi.org/10.1080/03075070903222633

Marquis, E., Fudge Schormans, A., Jung, B., Vietinghoff, C., Wilton, R., & Baptiste, S. (2016). Charting the Landscape of Accessible Education for Post-secondary Students with Disabilities. *Canadian Journal of Disability Studies, 5*(2), 42-71.

Marquis, E., Jung, B., Fudge-Schormans, A., Vajoczki, S., Wilton, R., Baptiste, S., & Joshi, A. (2012). Creating, Resisting or Neglecting Change: Exploring the Complexities of Accessible Education for Students with Disabilities. *The Canadian Journal for the Scholarship of Teaching and Learning, 3*(2).

Marquis, E., Jung, B., Fudge Schormans, A., Lukmanji, S., Wilton, R., & Baptiste, S. (2016). Developing inclusive educators: enhancing the accessibility of teaching and learning in higher education. *International Journal for Academic Development, 21*(4), 337-349.

Meleo-Erwin, Z., Kollia, B., Fera, J., Jahren, A., & Basch, C. (2021). Online support information for students with disabilities in colleges and universities during the COVID-19 pandemic. *Disability and Health Journal, 14*(1). doi:https://doi.org/10.1016/j.dhjo.2020.101013

Mendonça, C. R., Souza, K. T. O., Arruda, J. T., Noll, M., & Guimarães, N. N. (2021). Human Anatomy: Teaching-Learning Experience of a Support Teacher and a Student with Low Vision and Blindness. *Anat Sci Educ., Feb. 2*. doi:10.1002/ase.2058

Pacheco, L. F., Noll, M., & Mendonça, C. R. (2020). Challenges in teaching human anatomy to students with intellectual disabilities during the Covid-19 pandemic. *Anat Sci Educ., 13*(5), 556– 557. doi:10.1002/ase.1991

Pichette, J., Brumwell, S., & Rizk, J. (2020). *Improving the Accessibility of Remote Higher Education: Lessons from the Pandemic and Recommendations*: Higher Education Quality Council of Ontario.

Rose, D. H., & Meyer, A. (2002). *Teaching Every Student in the Digital Age: Universal Design for Learning*. Alexandria, VA.: Association for Supervision and Curriculum Development.

Scott, S., & Aquino, K. C. (2021). *COVID-19 Transitions: An Update on Access, Barriers, and Supports Nine Months into the Pandemic*: Association on Higher Education and Disability.

Snyder, T. D., de Brey, C., & Dillow, S. A. (2019). *Digest of Education Statistics 2018 (NCES 2020-009)*. Washington, DC.: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Soria, K. M., Horgos, B., Chirikov, I., & Jones-White, D. (2020). *The Experiences of Undergraduate Students with Physical, Learning, Neurodevelopmental, and Cognitive Disabilities During the Pandemic*. Retrieved from https://hdl.handle.net/11299/216715

UN. (2020). *Policy Brief: Education during COVID-19 and beyond*. Retrieved from https://www.un.org/development/desa/dspd/wp-content/uploads/sites/22/2020/08/sg_policy_brief_covid-19_and_education_august_2020.pdf

Zhang, H., Nurius, P., Sefidgar, Y., Morris, M., Balasubramanian, S., Brown, J., . . . Mankoff, J. (2021). How Does COVID-19 impact Students with Disabilities/Health Concerns? Retrieved from https://arxiv.org/abs/2005.05438

From Teaching to Teacher Training: Embedding Important Skills Needed to Develop a Teacher Trainer in Cascaded Teacher Professional Development Programmes

Lucian Vumilia NGEZE* & Sridhar IYER
IDP in Educational Technology, Indian Institute of Technology - Bombay, India
*lucianngeze@iitb.ac.in

Abstract: Cascaded teacher professional development (TPD) programmes can train many teachers within a short time frame compared to other models of teacher training. This is because trainers are teachers. Implementation of cascaded teacher training has been faced with many challenges as it puts emphasis on content knowledge. Trainers at different levels of the cascade model differ in terms of knowledge and skills. While many cascaded TPD programmes aim at creating secondary trainers, they are always trapped into transferring content knowledge alone to trainees. Skills are also needed to ensure they can transfer the learning to their contexts. This study used design based implementation research (DBIR) methodology in the design, development and implementation of the workshops that aimed to impart skills teachers need to become effective teacher trainers. Two content workshops with school teachers as participants were conducted, each followed by one skills workshop with 11 and 4 teacher trainers and experiences shared. The teacher trainers were selected based on set criteria at different stages. Analysis of open-ended data from teacher trainers showed that trainers mastered the skills and highlighted first steps when planning for a solo teacher training activity. This study contributes to the teacher training developers because, apart from the knowledge about the content, effective teacher trainers need different skills including skills on how to plan, conduct and evaluate teacher training workshops, participants and their contexts; development of activities and training materials, while preparing for a solo training.

Keywords: Cascade model, teacher trainers, secondary trainers, skills workshops, teacher professional development

1. Introduction

Teacher professional development (TPD) has been defined as activities that develop an individual knowledge, skills and attitudes as a teacher as they improve their teaching practices (Schleicher, 2009). Researchers emphasize that TPD activity should lead to improved knowledge in the domain area and teaching practices (Antoniou & Kyriakides, 2013).

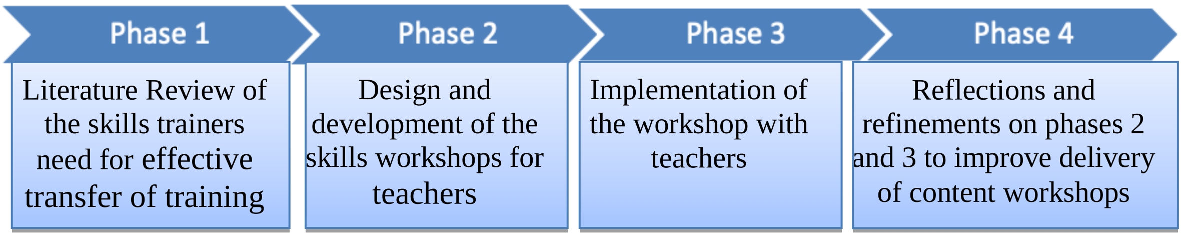
Effective TPD involves preparations in terms of training materials, process and trainers. While there are many approaches to TPD, workshops and seminars are the dominant modes of TPD in Tanzania. Teacher training workshops are faced with challenges such as too much content within a short time; very little time for teachers to reflect and focus more on content knowledge (Desimone, 2009). Even for practical based workshops, there is still a challenge of time to practice for the purpose of gaining deeper understanding of the content.

Many TPD activities in Tanzania are mainly planned and executed following a cascade model. Cascaded TPDs are associated with different challenges during implementation. These challenges can be categorized as design-related and trainer-related. From design perspective, cascaded TPDs have a top-down design approach (Komba & Mwakabenga, 2020); longer periods between cascade levels and one-way transmission. Literature has hinted out some of the trainer-related challenges as lack of confidence (Engelbrecht et al., 2007); curriculum misinterpretation (Suzuki, 2011); and dilution of the teaching content (Hayes, 2000).

This research focused on developing teacher trainers, looking into the skills needed to be able to successfully plan, conduct and evaluate a teacher training program. This study focused on finding answers to the following research questions (RQs).
RQ1: What personal skills do school teachers need to develop into school teacher trainers?
RQ2: How do teachers reflect about their past teacher training experiences after the skills workshop?
RQ3: How do teacher trainers prepare for solo training workshops?

2. Research Methodology

This research followed a Design Based Implementation Research (DBIR) methodology (Fishman et al., 2013). This methodology is suitable because it involves a series refinements and iterations to improve the intervention to the problem being solved. Figure 1 shows the different phases of DBIR.



Phase 1
Literature Review of the skills trainers need for effective transfer of training
Phase 2
Design and development of the skills workshops for teachers
Phase 3
Implementation of the workshop with teachers
Phase 4
Reflections and refinements on phases 2 and 3 to improve delivery of content workshops

Figure 1. DBIR Phases

2.1 Phase 1: The Skills Needed

Literature has highlighted the different skills any trainer needs to be effective. Experience in the area and facilitation skills are key to successful trainers' sessions (Ng & Lam, 2015). Another important aspect for teacher trainers is communication skills (Leach, 1996). This goes hand in hand with listening skills that are required when listening to the trainees' questions, presentations, and provide effective, timely and constructive feedback (Stolovich, 1999). To ensure that we imparted skills to the trainees who were transitioning to become teacher trainers, we developed two skills-based workshops.

2.2 Phases 2 and 3: Design, Development and Implementation of the Skills Workshops

Two skills were designed and developed based on the Attain Align Integrate (A2I) model (Warriem et al., 2014). These workshops were implemented in schools with teachers as participants. The details for each of the two workshops are discussed in the next subsection.

2.3 Phases 4: Reflections and Refinements on Phases 2 and 4

At the end of Skills Workshop 1, an evaluation study was conducted to determine areas that would need improvement and refinement for the next content workshop. These refinements led to iteration for Content Workshop 2 (CW2). In this way, the model for developing teacher trainers got improved and became more generalizable.

3. Skills Workshop Details and Data Collected

3.1 Skills Workshop 1 (SW1)

SW1 workshop involved 11 secondary school teachers all from different regions to Morogoro region in Tanzania. The 11 teachers had participated in the Content Workshop 1 (CW1). CW1 is a teacher workshop that aimed at training teachers on selected modules on ICT integration in teaching and learning. 26 teachers participated in CW1. Figure 2 shows the different content workshops (in yellow outline) and skills workshops (in blue). The number of participants per workshop is shown.

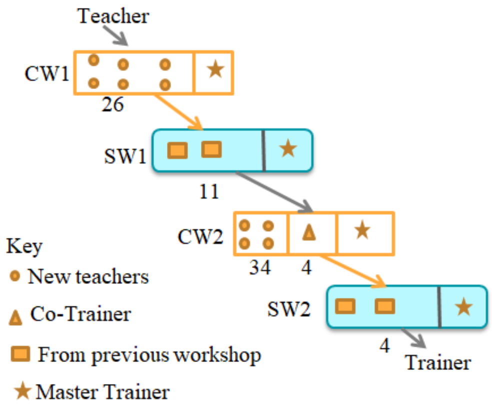


Figure 2. Content ans Skills Workshops.

To participate in SW1, participants met the following conditions: i) They had participated in CW1; ii) They should have applied what they gained in CW1 in their schools either through teaching his/her subject or sharing experience with other teachers; iii) Intrinsic motivation to participate in the skills workshop and iv) knowledge of some ICT tools. 11 teachers were selected to SW1 and were all males, with over five years of teaching experience. While most of the participants had some teacher training experience, only two have not had a chance to train teachers before.

SW1 was blended in such a way that, participants had first to complete four online activities (created in Moodle learning management system) that would make them join the face to face workshop a month later in Morogoro. Table 1 shows the online activities to impart the intended skills.

Table 1: *Online Activities for Imparting Teacher Training Skills*

Skill	Activity
Conducting needs analysis	Activity 1: Summarizing the details of the previous participants from Workshop Entry Survey
Contextualizing workshop content	Activity 2: Given sample topics to be covered during the workshop. Learners to go through the workshop materials and adapt them for their sessions
Question forming skills	Activity 3: Given a PI question and made to decide whether the given PI question is an effective question
Observation skills	Activity 4: Given a video case of one of the previous sessions of teacher workshops and then to observe the actions happening during the sessions

Face to face workshop session was conducted for two days in Morogoro, in December 2019. Day 1 involved a recap of the online activities submitted by participants. We also discussed on adult learning principles and how to incorporate them in the training. Day 2 focused on how to plan, conduct and evaluate teacher training programmes, including training requirements, managing training sessions and the Kirkpatrick model of training evaluation.

3.2 Skills Workshop 2 (SW2)

After SW1, selection criteria were set to select those who would become co-trainers in the next content workshop. The selection criteria included: participation in the previous skills workshop; confidence to train a small teacher training session in the next workshop; and availability for the next workshop in another new location (Mwanza). Only four participants met the criteria and were available for the next Content Workshop 2 (CW2). CW2 focused on helping the trainees from SW1 to apply all they had learnt. CW2 consisted of the same set of topics from CW1 but conducted to a new group of teachers in a different context (in Mwanza region). In CW2, the four trainees participated as co-trainers, working with the master trainer to ensure effectiveness in conducting the workshop. Before the workshop, the four co-trainers, each, were given a topic to prepare and train for duration not more than two hours. The goal was to apply the skills obtained in SW1 to a new context. Analysis of the data from this co-trained workshop revealed that more skills were needed to ensure that the co-trainers could be able to take up and engage in the sessions at the level that would increase independence.

SW2 focused on imparting more skills to the co-trainers to increase more independence when planning, conducting and evaluating teacher training workshops individually. This was a full online workshop that consisted of a one hour per day for 4 days. This involved the four co-trainers who participated in CW2. The topics emerged from reflections of participants in CW2 and are as listed below:

i. The training Cycle: To apply the steps while planning, conducting and evaluating the training
ii. Positive Climate in the Training Room: To manage the training room and emotion regulation
iii. Group Activities: To develop activity design skills, building confidence and presentation skills
iv. Evaluating the Training: Training evaluation skills.

3.3 Data Collection

For the study, different data collection methods and instruments were used. For each research question, one method was used to collect data. To know about the different skills teacher trainers need to become teacher trainers, we did a literature search. We used the following keywords: trainer skills, effective trainers, personal trainer skills, and training skills development. The skills identified as useful to the teacher trainers were used in designing and developing SW1.

During SW1, an interview protocol with six open-ended questions all focused to the co-trainers was used to collect data on reflections from teachers who joined with some teacher training experiences. The question being asked here was "*What has changed in terms of your teacher training sessions you had before completing these workshops*?" Their responses were recorded.

At the end of SW2, Google form was used to collect data from participants on their preparations for their solo teacher training sessions in their own contexts. The given scenario reads as:
A school in Arusha region has invited me (Master Trainer) to offer a 3-days teacher training to all 37 teachers in that school during this coming holiday (from June 7-9, 2021). I am planning to involve you to be the main trainer. I will support you via Zoom. How will you go about this training workshop?

3.4 Data Analysis

To find out the different skills trainers have to improve their training sessions, we did a literature review search using different keywords. These skills were categorized and are discussed in the results section.

The six phases of thematic analysis (Braun & Clarke, 2006) were used to generate themes for the co-trainers previous experience with teacher training during SW1. The collected data were first read and re-read to immerse into the data. Coding was done by the researcher, and initial codes were then generated from the data to identify some features from the data. The different codes were then sorted to generate initial themes by combining some codes to create a more meaningful theme. The themes generated were then reviewed and categorized and to create final themes as shown in Figure 3.

For the open-ended responses from the four co-trainers, summative content analysis (Hsieh & Shannon, 2005) was used to capture the occurrences of specific training preparation keywords in the responses for data collected during SW2. The frequencies of the keywords with the speaker or source were then identified. Finally, meaning and implication of the keywords were explained.

4. Results and Discussion

RQ1: What personal skills do school teachers need to develop into school teacher trainers?

Even though some co-trainers possessed a number of skills based on their experiences, workshop participants could identify some missing skills that co-trainers needed in order to improve their sessions. Workshop participants identified high voice, confidence and presentation skills as missing from the co-trainers sessions. Again, literature has highlighted different skills that trainers in teacher training programmes need to possess. Knowledge of participants and their contexts, question formation skills (Ng & Lam, 2015), observation skills and pedagogical knowledge of the content area are important skills. Other relevant skills include presentation skills, conducting and evaluate training (Leach, 1996). Different skills are needed to create an effective trainer.

RQ2: How do teachers reflect about their past teacher training experiences after the skills workshop?

Figure 3 depicts the themes generated from co-trainers experiences with the previous training sessions in their own contexts. It is clear that, the skills workshops have made a change in the ways co-trainers used to train in their schools. Training evaluation is one of the components for a successful training as hinted by (Aypay, 2009). The statements give a sense of a change that they will apply when they start planning for their individual teacher training workshops in their schools or beyond.

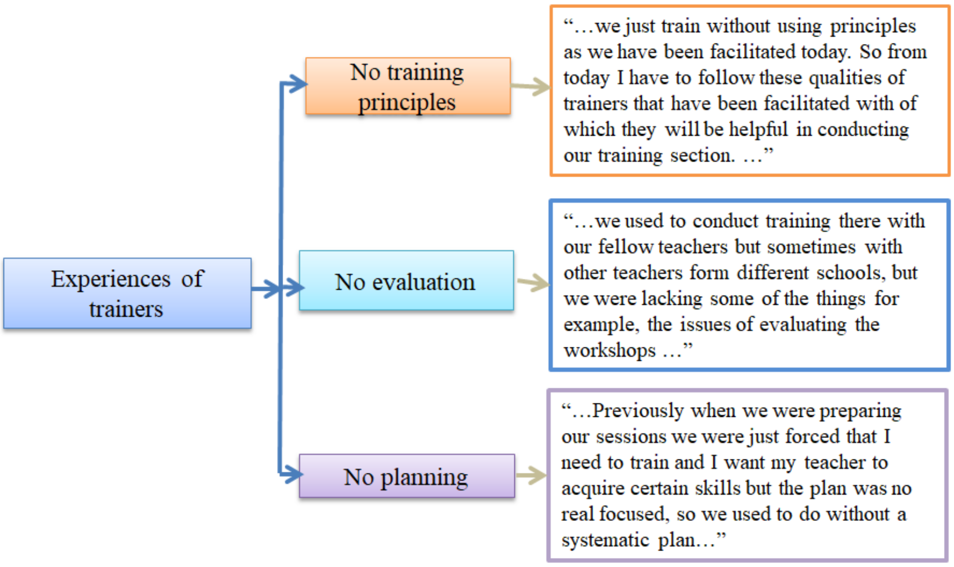

Experiences of trainers
No training principles
"…we just train without using principles as we have been facilitated today. So from today I have to follow these qualities of trainers that have been facilitated with of which they will be helpful in conducting our training section. …"
No evaluation
"…we used to conduct training there with our fellow teachers but sometimes with other teachers form different schools, but we were lacking some of the things for example, the issues of evaluating the workshops …"
No planning
"…Previously when we were preparing our sessions we were just forced that I need to train and I want my teacher to acquire certain skills but the plan was no real focused, so we used to do without a systematic plan…"

Figure 3. Themes on Experiences of Teacher Trainers.

RQ3: How do teacher trainers prepare for solo training workshops?

Planning, conducting and evaluating a training workshop is an important component in managing a teacher training. Even though the keywords are all self-preparation, support from the mentor is important (Feiman-Nemser, 2012). When asked about the preparations trainers would make to conduct solo training programmes, a number of keywords were mentioned, as shown in Table 2.

Table 2. *Important Codes Generated by Trainers*

Keywords/Codes	Frequency
Participants' needs analysis	4
Training action plan	3
Material preparation	4
Development of learning activities	3
Training evaluation	3

5. Research Implications

Successful cascaded teacher training programmes need secondary trainers who are experienced in teacher training. The first level of experience is by giving the trainee a chance to co-train the same session. This gives him/her one level of familiarity with the content and hence avoiding problems of misinterpretation of the content and dilution of the workshop content. On top of that, trainers have to go through a number of content and skill workshops that will make them competent to deliver a teacher training programme. The number of solo training workshops at one level will depend on the number of trainers made after CW2. Participants from the first solo training session become learners who then change their roles to co-trainers and them trainers in the preceding workshops. In this way the training programme becomes sustainable. Going through several workshops ensure content mastery to reduce dilution and distortion of the content. While certain skills might be highlighted in the training schedule, a follow up needs to be done to ensure the skills are being implemented during the training sessions. More skill practice time via different workshops is important to the development of a teacher trainer.

6. Conclusion

Teacher trainers need to be trained on different skills to be effective to train other teachers. Challenges of cascaded teacher training program such as lack of confidence to train others as the training moves to lower level of the cascaded TPD is minimized or reduced. Other challenges such as dilution of the training content and misinterpretation of the content are also minimized or removed completely as the trainers go through the same workshop content in CW1 and CW2. Based on the skills workshops conducted to teachers, this research suggests that trainer development is a process that needs time and skills to handle the training sessions.

Acknowledgements

We thank the school teachers from different regions of Tanzania who participated in these ICT workshops at different times. Special thanks to the four trainers who managed to participate in all CWs and SWs. Their participation and cooperation made this study possible.

References

Antoniou, P., & Kyriakides, L. (2013). A Dynamic Integrated Approach to teacher professional development: Impact and sustainability of the effects on improving teacher behaviour and student outcomes. *Teaching and Teacher Education*, *29*(1), 1–12. https://doi.org/10.1016/j.tate.2012.08.001

Aypay, A. (2009). Teachers' Evaluation of Their Pre-Service Teacher Training, Educational Sciences: Theory and Practice. *Educational Sciences: Theory and Practice*, *9*(3), 1113–1123.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38*(3), 181–199.

Dichaba, M. M. (2013). The perspectives of in-service trainers on the challenges of the Cascade model. *Anthropologist*, *15*(3), 265–275. https://doi.org/10.1080/09720073.2013.11891317

Engelbrecht, W., Ankiewicz, P., & De Swardt, E. (2007). An industry-sponsored, school-focused model for continuing professional development of technology teachers. *South African Journal of Education*, *27*(4), 579–595. https://doi.org/10.4314/saje.v27i4.25134

Feiman-Nemser, S. (2012). Beyond Solo Teaching. *Supporting Beginning Teachers*, *69*(8), 10–16.

Fishman, B., Penuel, W., Allen, A., Cheng, B., & Sabelli, N. (2013). : *Design-Based Implementation Research: An Emerging Model for Transforming the Relationship of Research and Practice*. *115*, 136–156.

Hayes, D. (2000). Cascade training and teachers' professional development. *ELT Journal*, *54*(2), 135–145.

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Komba, C., & Mwakabenga, R. (2020). Teacher Professional Development in Tanzania: Challenges and Opportunities. In *Educational Leadership*. IntechOpen. https://doi.org/10.5772/intechopen.90564

Ng, R., & Lam, R. (2015). Train-the-Trainer: A Study of the Professional Skill Competencies and Psychological Qualities of Teacher Trainer. *International Journal of Learning and Teaching*, *1*(1).

Schleicher, A. (2009). *Education at a Glance 2009: OECD Indicators. Report prepared for the Organization for Economic Co-Operation and Development*. http://hdl.voced.edu.au/10707/81434

Stolovich, H. (1999). Adult learning workshop. *Training 1999 Conference*.

Suzuki, T. (2011). *Cascade model for teacher training in Nepal*.

Warriem, J., Murthy, S., & Iyer, S. (2014). A2I: A Model for Teacher Training in Constructive Alignment for Use of ICT in Engineering Education. *In Proceedings of 22nd International Conference on Computers in Education*.

Suggestions for Special Education Teachers to Practice Spherical Image-based Virtual Reality Instruction in Classrooms: A Case Study

Kun-Hung CHENG*
Graduate Institute of Library and Information Science, National Chung Hsing University, Taiwan
*khcheng@dragon.nchu.edu.tw

Abstract: The benefits of learning by spherical image-based virtual reality (VR) have been reported in the field of formal education. However, there was limited VR research in practice for the instruction in special education. By implementing a series of spherical image-based VR instruction for English vocabulary learning in a special education class (i.e., participants including an experienced teacher and five students with cerebral palsy), this study aimed to provide instructional suggestions for special education teachers. Based on the qualitative results by thematic analysis of the teacher's interview transcripts and the researcher's field observation, this study discussed the practical issues for the instruction with the aid of spherical image-based VR technology for special education, such as learning process monitoring, adaptive worksheets design, incorporation of co-teachers, and adoption of VR assets for learning in daily situations.

Keywords: Virtual reality, special education, teacher, instruction

1. Introduction

Previous studies have documented the benefits of learning by virtual reality (VR) (e.g., enhancement of motivation, engagement, and experiential learning) with its capability of immersing students in situational learning contexts and facilitating interactivity between learners and computers (Dalgarno & Lee, 2010). However, these results were mostly generated in formal education environments. Scarce research explored the application of VR on special education, for example, teaching adaptive skills for adults with autism by VR (Schmidt et al., 2019). Therefore, there was limited VR research in practice for special education teachers to adopt in classrooms.

Considering the increasingly educational applications of immersive VR presenting spherical image-based or video-based virtual materials through low cost headsets such as Cardboard (e.g., Cheng & Tsai, 2019; Wu et al., 2019), there is a need to explore the affordance of low-cost VR in special education. Since a recent study (Geng et al., 2019) suggested the necessity to probe pedagogical use of VR in practice for teacher professional development, this study aimed to implement spherical image-based virtual reality instruction in special education fields and further provided instructional suggestions for special education teachers.

2. Method

This study invited a special education experienced teacher to implement a series of spherical image-based VR instruction for English learning in her class. Five eighth grade students (two females and three males) with different disability and handicaps of cerebral palsy in a special education school participated in this study. This study adopted an VR system, namely *Google Expeditions*, to immerse the students in spherical image-based virtual learning contexts. This system allows an instructor playing the role of a "tour guide" (using tablet PCs) to take his/her students playing the role of "tourists" (wearing VR headsets) to virtually visit some scenes such as historical sites or scenic spots with the presentation of 360° panoramic images (see Figure 1).

In this study, the two VR learning materials including *San Diego Zoo* (showing many animal exhibits and the natural environment in the San Diego Zoo) and *Underwater Galapagos* (presenting the underwater in the Galapagos Islands) in the *Google Expeditions* were adopted. Through observing the specific elements in the two virtual scenes such as sharks, sea lions, polar bears, or giraffes, the teacher guided her students to learn English vocabulary of animals.

The research trials consisted of three classes for three weeks (one class per week). In the first week, the researcher introduced how to setup the hardware and further demonstrated the virtual field trips in the classroom. The teacher implemented the VR learning activities in the second and third weeks. After the research trials finished, the teacher was interviewed in-depth for understanding her perceptions of the applications of spherical image-based VR learning for special education and the influences of VR technology on the students with cerebral palsy. The duration of interview was approximately 2 hours. In addition, the field observation on the teaching and learning in the classroom was conducted by the researcher.

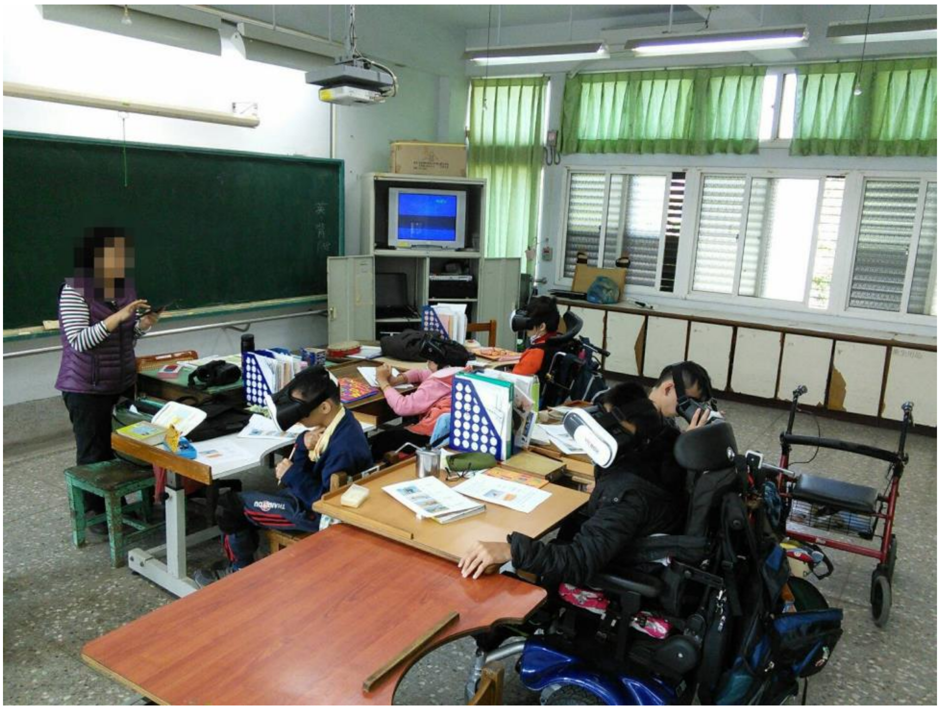


Figure 1. The Implementation of Virtual Field Trips in the Classroom

3. Results and Discussion

Through the thematic analysis of the interview transcripts of the teacher, there were several themes yielded. In general, the teacher expressed highly positive attitudes toward immersive VR learning for special education, particularly for enhancing the students' perceived presence and learning motivation. Based on these findings of the themes and the researcher's field observation, this study discussed the practical issues when implementing spherical image-based VR learning activities for special education.

- Teachers could monitor students' learning process according to individual differences through the functions of the *Google Expeditions* for managing their head tracking.
- Designing adaptive worksheets for English vocabulary learning based on the scaffolding information of the VR scenes provided by the system could be helpful for the evaluation of learning effectiveness.
- It was suggested to incorporate co-teachers to assist the students with different disability (e.g., poor coordination or stiff muscles) to adequately use the VR headsets.
- With the affordance of VR for the senses of presence, there were opportunities for special education teachers to adopt appropriate spherical image-based VR assets for the practice of English conversation in daily situations.

- In addition to domain knowledge learning, applying immersive VR learning for the training of life skills should be more important for special education.

References

Cheng, K. H., & Tsai, C. C. (2019). A case study of immersive virtual field trips in an elementary classroom: Students' learning experience and teacher-student interaction behaviors. *Computers & Education, 140,* 103600.

Dalgarno, B. D., & Lee, M. J. (2010). What are the learning affordances of 3-D virtual environments? *British Journal of Educational Technology, 41*(1), 10-32.

Geng, J., Chai, C. S., Jong, M. S. U., & Luk, E. T. H. (2019). Understanding the pedagogical potential of interactive spherical video-based virtual reality from the teachers' perspective through the ACE framework. *Interactive Learning Environments*, Advanced online publication.

Schmidt, M., Schmidt, C., Glaser, N., Beck, D., Lim, M., & Palmer, H. (2019). Evaluation of a spherical video-based virtual reality intervention designed to teach adaptive skills for adults with autism: a preliminary report. *Interactive Learning Environments*, Advanced online publication.

Wu, J., Guo, R., Wang, Z., & Zeng, R. (2019). Integrating spherical video-based virtual reality into elementary school students' scientific inquiry instruction: effects on their problem-solving performance. *Interactive Learning Environments*, Advanced online publication.

Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Computational Fluency and the Digital Divide in Japanese Higher Education

Luc GOUGEON & Jeffrey S. CROSS*
School of Environment and Society, Department of Transdisciplinary Science and Engineering, Tokyo Institute of Technology, Tokyo, Japan
*cross.j.aa@m.titech.ac.jp

Abstract: The COVID-19 pandemic has revealed a widening digital divide within the Japanese educational system. Most research on the digital divide focuses on K-12 education, whereas our research explores the situation more specifically in a Japanese higher education context. In higher education, we have identified a knowledge gap which can affect the career path of these students. We conducted a survey in 6 universities in the Kansai and Chugoku areas in Japan to map out the digital divide with different types of universities. This poster is intended to encourage stakeholder discussion in Japan to create a new curriculum for all higher education students and encourage all educators to pursue lifelong learning to improve their ICT and computational literacy skills.

Keywords: ICT, computational fluency, digital divide, Japanese education, higher education

1. Introduction

Japan's ambitions for the 21st century revolves around establishing what is termed as "Society 5.0," which aims at integrating advances in infrastructure, financial technology, healthcare, logistics and AI into the daily life of Japanese citizens (Minevich, 2019; Cabinet office, 2021) and improving the level of AI research before 2030 (Yamashita, 2019).

Japan faces a challenge when it comes to improving Information and Communication Technology (ICT) and computational fluency skills of its students and workers. Computational fluency was defined by Mitchel Resnick knowledge which is not limited to *computational concepts and problem-solving strategies, but also the ability to create and express oneself with digital technologies* (Resnick, 2018). Japan is currently ranking below the world average and often last in several international surveys such as the Programme for International Student Assessment (PISA,2020) 2018 (OECD, 2018) and Teaching and Learning International Survey (TALIS) (TALIS, 2020) in integration of ICT in education. This poster examines the computational fluency level of Japanese university students in order to identify their strengths and weaknesses in acquiring 21st century skills and attempts to answer the following questions:

- What is the ICT and computational literacy level of Japanese university students in the Kansai and Chugoku areas?
- Are there differences in the level of ICT and computational literacy between different private and national university students that were surveyed?

2. Method

A survey was distributed online to gather data about Japanese university students' computational fluency level. The survey was distributed from May 2020 to six higher education institutions.

Several questions of the survey were adapted from the International Computer and Information Literacy Study (ICILS) conducted by the International Association for the Evaluation of Educational Achievement (IEA). The ICILS is a measuring instrument aimed at 14 years old students, but some of the survey items are also applicable to adult students. The survey contains 55 questions pertaining to socio-economic background, computational fluency, and information literacy. We conducted a Wilcoxon rank sum test with continuity correction to identify significance.

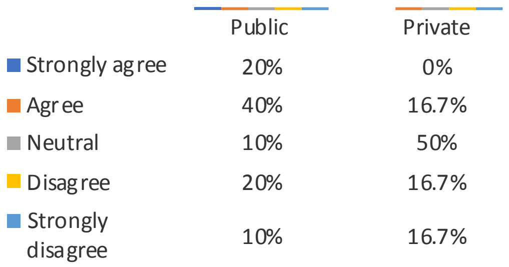
The survey received 481 responses from 6 national and private universities located in Hiroshima, Okayama, Hyogo, Osaka and Nara. The data is almost evenly split with 50.6% of the students at universities that are public and 49.4% being private

The *hensachi* scale is a numeric which is correlated with difficulty of admission into a university and is also a value used to rank universities (Makino, 2016; Toshin, 2021). National universities ranked between 50-60 and three out four private universities between 35 to 40 except for one private university in the Kansai area which ranked around 50, which is considered average. National universities in Japan are considered more prestigious than private universities, have lower tuition, and are more competitive to enter.

3. Results

Based upon the survey results, 36 out of 481 students reported not owning their own personal computer. 69% of the students who do not own a computer are studying in private universities The self-evaluation of ICT skills is illustrated in Table 1 reveals that students from public universities have higher comfort in using ICT than their counterpart in the private sector. Questions related to computer skills such as coding lessons prior to entering university reveals that 81% of students had never studied coding.

Table 1. *Students Comfort Level using ICT on A 5-Point Likert Scale*



	Public	Private
■ Strongly agree	20%	0%
■ Agree	40%	16.7%
■ Neutral	10%	50%
■ Disagree	20%	16.7%
■ Strongly disagree	10%	16.7%

The survey also collected basic socio-economic information about the students. The most relevant data is the highest level of education attained by the family of the student which is 9% higher from parents of students attending public universities than parents who send their children to private universities.

4. Discussion

Digital divide is defined as a growing gap between groups within a society regarding access to computers and the internet. Matsuoka is directly correlated to the education level of the family of the student with the level of inequality within the traditional educational system (Chokuron, 2021). 50.8 % of students from public universities come from household where parents hold a higher degree education while 41.8% of private universities students family hold a higher degree education. These socio-economic factors seem to be at the root of the digital divide within higher education in Japan.

A better understanding of the academic level of the high school that students attended before entering university and the data on cramming school attendance would have helped in highlighting the socio-economic factor underlying their ICT competence. Students from public universities or higher “hensachi” institutions considered themselves more comfortable in using ICT. None of the students from private universities felt very confident with the statement: “I am comfortable using ICT” while 20% of their counterpart in the public universities felt very confident. The current data does not allow us to clearly explain why students feel more confident in using ICT. The Wilcoxon rank sum test with continuity correction on confidence does not show significance with a p-value = 0.6618.

Task based measurement of computational fluency are needed to measure more accurately the level of students from private and public universities. The self-reporting of ICT confidence level would benefit from task-based assessment.

5. Conclusion

We have observed a digital knowledge gap between the private and public institutions surveyed in the Kansai and Chugoku area which indicates the presence of a digital divide in higher education. This digital divide is added to the advantages that students from higher level universities already have over students from lower-level universities. Japan can reverse this trend and position itself in line with the SDG Goal number 4 by improving digital literacy and access to educational technology to all its students at all levels.

The public k-12 educational system and higher education institutions have the responsibility to include computational fluency within their curriculum. Failure to do so, would favor the students who can afford private lessons known as cramming school in Japan.

This disparity in skills can affect students' chances in finding work after graduation can be addressed by improving the ICT skills of all Japanese students. By adopting new educational policies aimed at improving computational fluency within the public educational sector, the Japanese government can reduce the digital divide within higher education. The COVID-19 pandemic has prompted the Japanese government to accelerate the digitalization of all institutions and this effort will ne to be supported with training in computational fluence at scale.

6. Future work

We also intend to further study the digital divide and socio-economic background of university students by conducting more in-depth surveys and present the results in the poster presentation.

References

Minevich, M. (2019) Japan's Society 5.0' initiative is a road map for today's entrepreneurs, *TechCrunch* Retrieved September 8, 2021, from https://techcrunch.com/2019/02/02//japans-society-5-0-initiative-is-a-roadmap-for-todays-entrepreneurs/

Cabinet office (2019) What is Society 5.0? *Cabinet Office, Government of Japan*. Retrieved September 8, 2021, from https://www8.cao.go.jp/cstp/english/society5_0/index.html

Okubo, T. (2021) Expansion of Disparities During the COVID-19 Pandemic - The Income Gap and the Digitalization Gap. *Nippon Institute for Research advancement*. No.53.

OECD. (2020). School Education During Covid 19, Were Teachers and Students Ready? *Japan country note. OECD*, Retrieved from https://oecd.org/education/Japan-coronavirus-education-country-note.pdf

TALIS - The OECD Teaching and Learning International Survey. (2020) *OECD*. Retrieved from https://oecd.org/edcation/talis/

Makino, M. (2016). Times Higher Education Rankings and Hensachi in Japanese Universities. *Mark: My words*, Retrieved September 8, 2021, from https://futurealisreal.wordpress.com/2016/10/09/times-higher-education-rankings-and-hensachi-in-

Chokuron. (2021).コロナ禍と教育格差：ICT活用後進国ニッポンの大問題. *Chuokoron Shinsha*, Retrieved September 8, 2021,from https://chuokoron.jp/society/114417.html

IEA. (2020). International Computer and Information Literacy Study. *IEA*. Retrieved September 8, 2021, from https://iea.nl/studies/iea/icils

Toshin (2021) Toshin Hensachi, Retrieved September 8, 2021, from https;//toshin-hensachi.com

McVeigh, B. (2002) Japanese Higher Education as Myth. *M.E. Sharpe*, New York.

Resnick, M. Computational Fluency. Retrieved September 8, 2021, from https://mres.medium.com/computational-fluency-776143c8d725

Data-informed Teaching Reflection: A Pilot of a Learning Analytics Workflow in Japanese High School

Taro NAKANISHI[a*], Hiroyuki KUROMIYA[a] , Rwitajit MAJUMDAR[b] & Hiroaki OGATA[b]
[a]*Graduate School of informatics, Kyoto University, Japan*
[b]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*nakanishi.taro.26r@st.kyoto-u.ac.jp

Abstract: The use of ICT devices in the classroom for teaching has become more relevant in secondary education. However, the use of data driven technologies for supporting pedagogical practices are still limited in the Japanese school context. Promoting the use of ICT in the classroom is an important part of improving the quality of education. In this paper, we introduce the workflow of using e-book reader and learning analytics dashboard to improve teaching of a junior-high school Mathematics class in Japan. We recruited a teacher and forty students for the research and introduced the simple dashboard for daily teaching and learning in their class. An end of the study period survey results showed that the dashboard prompted a change in the way of teaching in the class and also the teacher became more positive towards using data-driven technologies for reflecting on teaching practices.

Keywords: Learning analytics, evidence-based education, secondary education, ebook reader, LEAF

1. Introduction

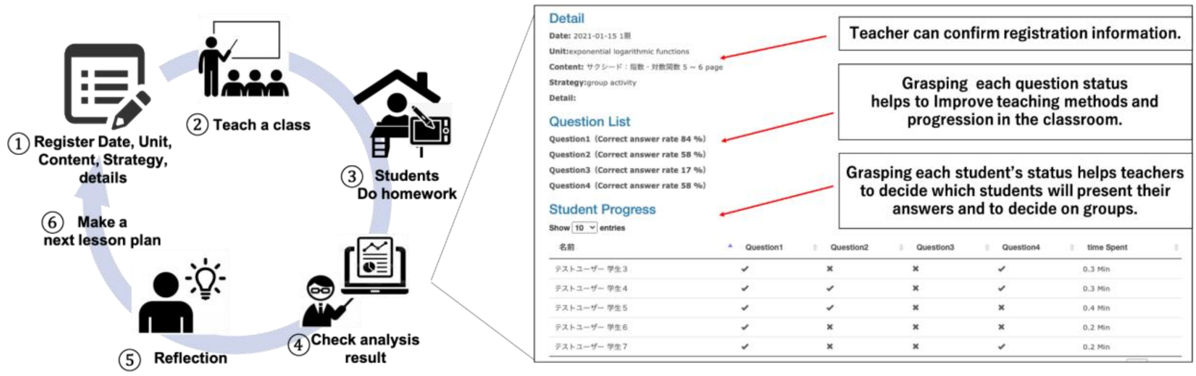
In recent years, the use of ICT in education has become more popular around the world (Majumdar, 2015). This adoption comes with numerous benefits. For example, the use of ICT promotes positive motivation (Harandi, 2015), teachers can also teach proactively and cooperatively (Aristovnik, 2012). Moreover, it leads to achieving individual learning and improving the quality of education. On the other hand, it is known that the effectiveness of ICT depends on the ability of teachers (Hernandez, 2017). However, one study (Tasaki, 2017) found, Japanese teachers have the lowest rate of use of digital devices among OECD countries . One solution to this problem is to promote teachers' skills in using ICT in the classroom and to develop a positive attitude towards the value of ICT in education (Gil-Flores et al., 2017). On the other hand there are benefits of data driven approaches in education (Tempelaar et al., 2021) which can be realised once the teacher adopts the practices.

The main aim of this research is to investigate the effect of data-driven teacher reflection supported by a learning analytics (LA) technology framework. In this paper, we propose a workflow for teachers to improve their teaching based on data provided in a LA dashboard with a case study of a high school mathematics teacher in Japan. This pilot observation was guided by the following research question: How does the proposed teachers' workflow affect practice in a regular classroom context?

2. Methods

2.1 Teachers' Workflow to Improve Practice

The workflow consists of 6 different steps (see Figure 1.). The teachers register the date, unit, content, strategy, details in the system and then proceed to teach the class. To collect the data, the students do their homework on the tablet. Teachers then check the analysis results of the homework through the system. This is followed by a reflection on the class about the level of the problems covered in the lesson and homework, the difference between the teacher's expectations and the students' level of understanding, and points for improvement. Finally, based on reflection, the teachers make the next lesson plan. There are two advantages of this workflow. First, the system can help to aggregate the students interactions and responses regarding the homework tasks and presents that information in the dashboard. Teachers can focus on how to improve their teaching from the result. Second, they can check the results of the improvements in each lesson. Therefore, they can adjust their lesson plans and teaching to meet the understanding of their students.



① Register Date, Unit, Content, Strategy, details
② Teach a class
③ Students Do homework
④ Check analysis result
⑤ Reflection
⑥ Make a next lesson plan

Detail

Date: 2021-01-15 1限

Unit:exponential logarithmic functions

Content: サクシード：指数・対数関数 5 ~ 6 page

Strategy:group activity

Detail:

Question List

Question1 (Correct answer rate 84 %)

Question2 (Correct answer rate 58 %)

Question3 (Correct answer rate 17 %)

Question4 (Correct answer rate 58 %)

Student Progress

Show 10 entries

名前	Question1	Question2	Question3	Question4	time Spent
テストユーザー 学生3	✓	✗	✗	✓	0.3 Min
テストユーザー 学生4	✓	✓	✗	✓	0.3 Min
テストユーザー 学生5	✓	✓	✗	✗	0.4 Min
テストユーザー 学生6	✓	✗	✗	✗	0.2 Min
テストユーザー 学生7	✓	✗	✗	✓	0.2 Min

Teacher can confirm registration information.

Grasping each question status helps to Improve teaching methods and progression in the classroom.

Grasping each student's status helps teachers to decide which students will present their answers and to decide on groups.

Figure 1. Teachers Workflow to Improve Practice.

2.2 Technical Design

In our research, we use an e-book system, BookRoll (Ogata et al., 2018) to enable students and teachers to view and interact with the learning materials. Teachers can upload teaching material in pdf format and make quizzes on BookRoll. Students can access the BookRoll system through a standard web browser. When they interact with a page in a learning material (move page, click, add marker, add memo, answer quizzes etc.), the system monitors and collects log data.

The Analysis Tool (Akçapınar et al., 2019) analyzes the log data from the database and displays the results on the dashboard. We developed and integrated a class planning system on Analysis Tool. This system has two functions. First, teachers can register class information (date, unit, content, strategy, details.) (for step 1 in Figure 1) . Second, after students view teaching materials and do their homework on BookRoll, the class planning system displays the response rate, percentage of correct answers, and average time to answer. The teachers can also view detailed information on individual students.

3. Initial Pilot in a High School Mathematics Class

This research involved the first grade Mathematics class at a high school in Japan. A teacher and 40 students from the first year high school class were involved in this research for 1 month. These participants have been using the tablet to learn on a daily basis for 1 year. The teacher was also accustomed to working with the mobile devices and used the ebook reader to upload teaching materials and check analysis results in the LAViEW dashboard. The students received instruction in a unit on exponential logarithmic functions over 10 lectures over a period of about one month.

3.1 Perceived and Observed Effects

A questionnaire was administered to the period before and after the demonstration. The results showed that the teacher was able to use concrete data to improve their teaching and to understand their teaching skills compared to before the demonstration. The teacher was particularly interested in the percentage of correct responses from each student when considering interventions to improve their teaching. Specifically, the teacher explained questions with low correct answers during class, and asked students who answered correctly to explain how to solve the question. The teacher wanted to improve the daily lesson content according to the students' level of understanding, but this had been difficult in the past due to time constraints. However, he responded that the new system allowed him to monitor each student’s level of understanding in a short period of time and that it was easier to make improvements because he could continue to see the level of understanding on a daily basis.

4. Conclusion

In this paper, we have presented a workflow for teachers to improve their teaching based on usage of learning analytics driven technology. The aim of the research was to design a data-driven teacher reflection workflow and support it by a learning analytics technology framework. The initial deployment and observation in one class showed changes in teachers' teaching style, such as changing questions according to student’s understanding and explaining questions with low percentages of correct answers over a month. In addition, teacher’s awareness of data-driven teaching increased and their positive attitude towards the use of ICT was also observed. However, these initial interactions with the teacher and students also let us realise some importanaspects. For instance quantitative evaluation of any improvement in teaching would require further studies and validated metrics for measurements. Also there needs refinement of how to introduce any such data-driven system and its workflow to the students as well as the teachers. The results of this research show that data-driven interventions by teachers can contribute to raising teacher’s awareness of ICT and to improve their teaching. Repeated data-driven interventions and monitoring of their effectiveness by teachers will enable them to improve both their ICT skills and their teaching skills. Our future work will involve supporting teachers to select effective interventions based on recommendation from the data.

Acknowledgements

This research is supported by JST JPM-JAX20AA, JSPS 21J14514, SPIRITS 2020 of Kyoto University, JSPS 20K20131, JSPS 20H01722, JSPS 16H06304, NEDO JPNP18013, and NEDO JPNP20006. We thank all the teachers and students in the target school for their kind cooperation during this research.

References

Akçapınar, G., Hasnine, M. N., Majumdar, R., Flanagan, B., & Ogata, H. (2019). Using learning analytics to detect off-task reading behaviors in class. In *Companion Proceedings of the 9th International Conference on Learning Analytics and Knowledge (LAK'19),* 471-476. Society for Learning Analytics Research (SoLAR).

Aristovnik, A. (2012). The Impact of ICT on Educational Performance and its Efficiency In Selected EU and OECD Countries: A Non-Parametric Analysis. *SSRN Electronic Journal*.

Gil-Flores, J., Rodríguez-Santero, J., & Torres-Gordillo, J. J. (2017). Factors that explain the use of ICT in secondary-education classrooms: The role of teacher characteristics and school infrastructure. *Computers in Human Behavior, 68*, 441–449.

Harandi, S. R. (2015). Effects of e-learning on Students’ Motivation. P*rocedia-Social and Behavioral Sciences, 181*, 423-430.

Hernandez, R. M. (2017). Impact of ICT on Education: Challenges and Perspectives. *Journal of Educational Psychology-Propositosy Representaciones, 5(1)*, 337-347.

Majumdar, S. (2015). Emerging trends in ICT for education & training. *Gen. Asia Pacific Reg. IVETA.*

Ogata, H., Majumdar, R., Akçapınar, G., Hasnine, M. N., & Flanagan, B. (2018). Beyond learning analytics: Framework for technology-enhanced evidence-based education and learning. In *procs. of 26th ICCE Workshop Proceedings*, 493-496.

Tasaki, N. (2017). The impact of OECD- PISA results on Japanese educational policy. *European Journal of Education*, 52(2), 145-153.

Tempelaar, D., Rienties, B., & Nguyen, Q. (2021). The contribution of dispositional learning analytics to precision education. *Educational Technology & Society, 24*(1), 109-122.