

# Journal of Educational Technology & Society

Published by International Forum of Educational Technology & Society  
Hosted by National Yunlin University of Science and Technology, Taiwan

Jul. 2021

Journal of **Educational Technology & Society**

The Journal of Educational Technology & Society has Impact Factor 3.522 and  
5-Year impact factor 3.941 according to Thomson Scientific 2020 Journal Citations Report.

<http://www.j-ets.net/>

vol. **24**  
no. **3**

Volume **24** Issue **3**

**July 2021**

ISSN: 1436-4522 (online)  
ISSN: 1176-3647 (print)  
<http://www.j-ets.net/>



# Educational Technology & Society

An International Journal

## Aims and Scope

The journal of *Educational Technology & Society* (ET&S) is an open-access academic journal published quarterly (January, April, July, and October) since October 1998. By 2018, ET&S has achieved its purposes at the first stage by providing an international forum for open access scientific dialogue for developers, educators and researchers to foster the development of research in educational technology. Thanks to all of the Authors, Reviewers and Readers, the journal has enjoyed tremendous success.

Starting from 2019, the ET&S journal has established a solid and stable editorial office with the support of National Yunlin University of Science and Technology. The new Editors-in-Chief have been appointed aiming to promote innovative educational technology research based on empirical inquiries to echo the pedagogical essentials of learning in the real world—lifelong learning, competency-orientation, and multimodal literacy in the 21st century.

ET&S publishes the research that well bridges the pedagogy and practice in advanced technology for evidence-based and meaningfully educational application. The focus of ET&S is not only technology per se, but rather issues related to the process continuum of learning, teaching, and assessment and how they are affected or enhanced using technologies rooted in a long-period base. The empirical research about how technology can be used to overcome the existing problems in the frontline of local education with findings that can be applied to the global spectrum is also welcome. However, papers with only descriptions of the results obtained from one hit-and-run and short-term study or those with the results obtained from self-report surveys without systematic or empirical data or any analysis on learning outcomes or processes are not favorable to be included in ET&S.

## Founding Editor

**Kinshuk**, University of North Texas, USA.

## Journal Steering Board

**Nian-Shing Chen**, National Yunlin University of Science and Technology, Taiwan; **Kinshuk**, University of North Texas, USA; **Demetrios G. Sampson**, University of Piraeus, Greece.

## Editors-in-Chief

**Maiga Chang**, Athabasca University, Canada; **Andreas Harrer**, Dortmund University of Applied Sciences and Arts, Germany; **Yu-Ju Lan**, National Taiwan Normal University, Taiwan; **Yu-Fen Yang**, National Yunlin University of Science and Technology, Taiwan.

## Editorial Board Members

**Ahmed Hosny Saleh Metwally**, Northeast Normal University, China; **Bernardo Pereira Nunes**, The Australian National University, Australia; **Ching-sing Chai**, The Chinese University of Hong Kong, Hong Kong; **David Gibson**, Curtin University, Australia; **Grace Yue Qi**, Massey University, New Zealand; **Ig Ibert Bittencourt Santana Pinto**, Universidade Federal de Alagoas, Brazil; **Jerry Chih-Yuan Sun**, National Chiao Tung University, Taiwan; **Jie Chi Yang**, National Central University, Taiwan; **Joice Lee Otsuka**, Federal University of São Carlos, Brazil; **Kaushal Kumar Bhagat**, Indian Institute of Technology, India; **Minhong Wang**, The University of Hong Kong; **Morris Siu-Yung Jong**, The Chinese University of Hong Kong; **Regina Kaplan-Rakowski**, University of North Texas, USA; **Rita Kuo**, New Mexico Tech, USA; **Robert Li-Wei Hsu**, National Kaohsiung University of Hospitality and Tourism, Taiwan; **Rustam Shadiev**, Nanjing Normal University, China; **Stephen J.H. Yang**, National Central University, Taiwan; **Tony Liao**, NOAA Earth System Research Laboratories, USA; **Wen-Ta Tseng**, National Taiwan University of Science and Technology, Taiwan; **Yanjie Song**, Education University of Hong Kong; **Yun Wen**, National Institute of Education, Singapore.

## Managing Editor

**Sie Wai (Sylvia) Chew**, National Sun Yat-sen University, Taiwan; **Phaik Imm (Alexis) Goh**, National Yunlin University of Science and Technology, Taiwan.

## Editorial Assistant

**Kao Chia-Ling Gupta**, The University of Hong Kong, China; **Yen-Ting R. Lin**, National Taiwan Normal University, Taiwan.

## Technical Manager

**Wei-Lun Chang**, National Yunlin University of Science and Technology, Taiwan.

## Executive Peer-Reviewers

see <http://www.j-ets.net>

## Publisher

International Forum of Educational Technology & Society

## Host

National Yunlin University of Science and Technology, Taiwan

## Editorial Office

c/o Chair Professor Nian-Shing Chen, National Yunlin University of Science and Technology, No. 123, Section 3, Daxue Road, Douliu City, Yunlin County, 64002, Taiwan.

## Supporting Organizations

University of North Texas, USA  
University of Piraeus, Greece

## Advertisements

*Educational Technology & Society* accepts advertisement of products and services of direct interest and usefulness to the readers of the journal, those involved in education and educational technology. Contact the editors at [journal.ets@gmail.com](mailto:journal.ets@gmail.com)

## Abstracting and Indexing

*Educational Technology & Society* is abstracted/indexed in Social Science Citation Index, Current Contents/Social & Behavioral Sciences, ISI Alerting Services, Social Scisearch, ACM Guide to Computing Literature, Australian DEST Register of Refereed Journals, Computing Reviews, DBLP, Educational Administration Abstracts, Educational Research Abstracts, Educational Technology Abstracts, Elsevier Bibliographic Databases, ERIC, JSTOR, Inspec, Technical Education & Training Abstracts, and VOCED.

## Guidelines for authors

Submissions are invited in the following categories:

- Peer reviewed publications: Full length articles (4,000 to 8,000 words)
- Special Issue publications

All peer review publications will be refereed in double-blind review process by at least two international reviewers with expertise in the relevant subject area.

For detailed information on how to format your submissions, please see:

[https://www.j-ets.net/author\\_guide](https://www.j-ets.net/author_guide)

For Special Issue Proposal submission, please see:

[https://www.j-ets.net/journal\\_info/special-issue-proposals](https://www.j-ets.net/journal_info/special-issue-proposals)

## Submission procedure

All submissions must be uploaded through our online management system (<http://www.j-ets.net>). Do note that all manuscripts must comply with requirements stated in the Authors Guidelines.

Authors, submitting articles for a particular special issue, should send their submissions directly to the appropriate Guest Editor. Guest Editors will advise the authors regarding submission procedure for the final version.

All submissions should be in electronic form. Authors will receive an email acknowledgement of their submission.

The preferred formats for submission are Word document, and not in any other word-processing or desktop-publishing formats. For figures, GIF and JPEG (JPG) are the preferred formats. **Authors must supply separate figures** in one of these formats besides embedding in text.

Please provide following details with each submission in a separate file (i.e., Title Page): ■ Author(s) full name(s) including title(s), ■ Name of corresponding author, ■ Job title(s), ■ Organisation(s), ■ Full contact details of ALL authors including email address, postal address, telephone and fax numbers.

In case of difficulties, please contact [journal.ets@gmail.com](mailto:journal.ets@gmail.com) (Subject: Submission for Educational Technology & Society journal).

## Table of contents

### Full Length Articles

- The Professionalism of Online Teaching in Arab Universities: Validation of Faculty Readiness 1–12  
*Ahmed Ramadan Khtere and Ahmed Mohamed Fahmy Yousef*
- Exploring Effects of Geometry Learning in Authentic Contexts Using Ubiquitous Geometry App 13–28  
*Wu-Yuin Hwang, Uun Hariyanti, Yan Amal Abdillah and Holly S. L. Chen*
- The Impact of Game Playing on Students' Reasoning Ability, Varying According to Their Cognitive Style 29–43  
*Tsung-Yen Chuang, Martin K.-C. Yeh and Yu-Lun Lin*
- Flipped Classroom in the Educational System: Trend or Effective Pedagogical Model Compared to Other Methodologies? 44–60  
*Héctor Galindo-Dominguez*
- Interaction Effects of Situational Context on the Acceptance Behaviour and the Conscientiousness Trait towards Intention to Adopt: Educational Technology Experience in Tertiary Accounting Education 61–84  
*Mohamad Ridhuan Mat Dangi and Maisarah Mohamed Saat*

### Editorial Note

- From Conventional AI to Modern AI in Education: Re-examining AI and Analytic Techniques for Teaching and Learning 85–88  
*Haoran Xie, Gwo-Jen Hwang and Tak-Lam Wong*

### Special Issue Articles

- Perceptions of and Behavioral Intentions towards Learning Artificial Intelligence in Primary School Students 89–101  
*Ching Sing Chai, Pei-Yi Lin, Morris Siu-Yung Jong, Yun Dai, Thomas K. F. Chiu and Jianjun Qin*
- Gender Differences in Cognitive Load when Applying Game-Based Learning with Intelligent Robots 102–115  
*Beyin Chen, Gwo-Haur Hwang and Shen-Hua Wang*
- Factors Affecting the Adoption of AI-Based Applications in Higher Education: An Analysis of Teachers' Perspectives Using Structural Equation Modeling 116–129  
*Youmei Wang, Chenchen Liu and Yun-Fang Tu*
- Prediction of Student Performance in Massive Open Online Courses Using Deep Learning System Based on Learning Behaviors 130–146  
*Chia-An Lee, Jian-Wei Tzeng, Nen-Fu Huang and Yu-Sheng Su*
- Automatic Generation of Cloze Items for Repeated Testing to Improve Reading Comprehension 147–158  
*Albert C. M. Yang, Irene Y. L. Chen, Brendan Flanagan and Hiroaki Ogata*
- Expert-Authored and Machine-Generated Short-Answer Questions for Assessing Students' Learning Performance 159–173  
*Owen H. T. Lu, Anna Y. Q. Huang, Danny C. L. Tsai and Stephen J. H. Yang*
- Effects of Personalized Intervention on Collaborative Knowledge Building, Group Performance, Socially Shared Metacognitive Regulation, and Cognitive Load in Computer-Supported Collaborative Learning 174–193  
*Lanqin Zheng, Lu Zhong, Jiayu Niu, Miaolang Long and Jiayi Zhao*



Teachable Agent Improves Affect Regulation: Evidence from Betty's Brain <i>Jian-Hua Han, Keith Shubeck, Geng-Hu Shi, Xiang-En Hu, Lei Yang, Li-Jia Wang, Wei Zhao, Qiang Jiang and Gautum Biswas</i>	194–209
Exploring the Relationships between Achievement Goals, Community Identification and Online Collaborative Reflection: A Deep Learning and Bayesian Approach <i>Changqin Huang, Xuemei Wu, Xizhe Wang, Tao He, Fan Jiang and Jianhui Yu</i>	210–223
STEM-based Artificial Intelligence Learning in General Education for Non-Engineering Undergraduate Students <i>Chun-Hung Lin, Chih-Chang Yu, Po-Kang Shih and Leon Yufeng Wu</i>	224–237
Progress, Challenges and Countermeasures of Adaptive Learning: A Systematic Review <i>Fengying Li, Yifeng He and Qingshui Xue</i>	238–255
A Bayesian Classification Network-based Learning Status Management System in an Intelligent Classroom <i>Chuang-Kai Chiu and Judy C. R. Tseng</i>	256–276

# The Professionalism of Online Teaching in Arab Universities: Validation of Faculty Readiness

Ahmed Ramadan Khater<sup>1\*</sup> and Ahmed Mohamed Fahmy Yousef<sup>2</sup>

<sup>1</sup>Foundations of Education Department, Faculty of Education, Fayoum University, Egypt // <sup>2</sup>Educational Technology Department, Faculty of Specific Education, Fayoum University, Egypt // ahmed.s.a@fayoum.edu.eg // ahmed.fahmy@fayoum.edu.eg

\*Corresponding author

(Submitted October 16, 2020; Revised December 20, 2020; Accepted February 20, 2021)

**ABSTRACT:** The study aimed to examine the readiness of faculty members in Arab universities for blended learning environments through an investigation of the attributes, skills, and knowledge in three roles of professional online teachers. Online teaching professionalism has been described as a set of required competencies, and behaviours for the effectiveness of educational online sessions. The authors have argued some requirements of teachers' roles as an instructional planner, an assessor, and as a mentor. A purposive sample of 24 experts from diverse disciplines contributed to the reference panel in a Delphi study through three rounds. Qualitative content analysis and some descriptive statistics e.g., the median and frequency distribution, have been used to reach a consensus among the panel of experts. A matrix of 30 requirements was shortlisted by experts in different roles. The panelists provided insight into the top 10 requirements for each role to measure the professionalism of faculty before, during, and after the online sessions. The readiness for online teaching was concluded by six main domains namely evaluating students' achievements and limitations, problem-solving skills, information technology and computer skills, monitoring and motivating techniques, communication, and class management skills. The study results can be used to plan faculty development programs based on performance gaps of faculty members at three levels: individual, departmental, and program or college. Moreover, the listed faculty attributes help higher education institutions to evaluate the perceptible skills and personal characteristics of faculty in enhancing the efficacy of online teaching in different academic disciplines.

**Keywords:** Teaching professionalism, Online learning, Faculty readiness, Faculty competences

## 1. Introduction

The educational system around the globe has been disrupted to varying degrees due to Covid-19 prevailing. The United Nations Educational, Scientific and Cultural Organization, "UNESCO," has counted that more than 1.5 billion students in 165 countries have been forced to drop out of schools and universities. The pandemic forced academic bodies around the world to discover new patterns of learning and education. In response to this threat, new ideas towards online learning strategies are emerging, being tested, and evaluated, albeit with a lot of effort and challenges for teachers and parents (UNESCO, 2020).

Considering this, we conducted 42 semi-structured interviews with faculty members from different disciplines in some Arab universities, with the aim of seeking their opinions regarding distance learning experience during Covid-19 lockdown. The most important finding was that the university staff developed their own teaching strategy for online classes. In fact, it differs from one individual to another, and from one university to another according to the circumstances and the available capabilities. Most of the instructors started preparing their educational materials electronically, without following certain standards, and 90% of them asserted the necessity to cover most aspects of the curriculum, after teaching online lectures, curriculum materials being uploaded to the learning management system (LMS). Furthermore, some Open Educational Resources (OER) were used by 25% of instructors and posted to the LMS portal for those who missed the classes due to some inevitable circumstances. The interviewees agreed that the instructor plays a major role in implementing online learning strategies, as he acts as a guide for students, a catalyst for them, and an instructional designer to use the technology through which learning takes place, and provide effective and constructive feedback, following up on the level of students and providing the necessary recommendations on time.

Thus, the professionalism of online teaching has become a relevant topic of discussion among educators and academics for continuing work in education and teaching in 2020. As the prevalence of blended learning and online courses in higher education institutions increases, so does the need for research on faculty competencies and skills in those online environments. Rapanta et al. (2020) argued that universities, to be competitive during and after the Covid-19 crisis, have to adopt some indications of faculty preparedness, in terms of their

professionalism, which is necessary for online teaching as an essential function of such professional preparedness. As such, research on the educational requirements, and the challenges of teaching in a diverse environment, is the current hot topic of the day with fundamental changes in some universities. Some researchers concluded that teachers want to explore ways to create a more engaging and effective environment for themselves and their students.

Teaching professionalism has been identified in many previous studies as a key element that permeates two standards of faculty competencies namely personal, and professional competencies. More specifically, teaching professionalism involves curriculum design, delivery, and oversight (Shelly & Scolaro, 2016). Hence, closer scrutiny of these competencies will provide a depth understanding of what faculty of online classes should have to reflect on their philosophy of teaching, make it crystal clear to students, and implement it steadily and explicitly. Indeed, many of the educational theories, such as social constructivism, connectivism, situated learning and communities of practice, have been exploring by educational theorists to investigate where and how can be used to enhance online learning (Ni She et al., 2019). The findings of exploring factors influencing faculty revealed three primary approaches to teaching online, namely content acquisition, collaborative learning, and knowledge building, which are relevant to some factors e.g., faculty age, their academic background, and online teaching dedication (Badia et al., 2017).

Given the above, higher education institutions need to examine the experiences that online educators face in a virtual setting, such as strengths, challenges, perceived level of professionalism, and perspectives on the future online teaching (Sims, 2017). Moreover, provide faculty the professional development which can develop their abilities to support the application of diverse and appropriate learning theories. Hence, supporting conceptual change should be a central constituent of professional development activities if more effective use of educational technology is to be achieved (Englund et al., 2017). This can be useful in terms of helping to recognize the methodological criteria which to a great extent guarantee the effectiveness of training in two perspectives: meeting faculty training needs and, consequently, improving teacher practices in university virtual environments (Alvarez et al., 2009). Higher education institutions need to frequently evaluate the challenges that faculty face in the design and delivery of courses through virtual learning environments, and to prioritize efforts to remediate them (Kibaru, 2018; Mishra et al., 2020).

Ideally, Delphi technique has been functional in higher education to evaluate and establish a communication structure aimed at constructing a comprehensive critical examination and discussion of instructional design principles, challenges in establishing adaptive learning, campus environment, and institutional research by using the constructs or canons of credibility and confirmability (Mirata et al., 2020; Green, 2014). Several Delphi studies have been used, in different academic disciplines, to recognize and develop the professional attributes. Rowe et al. (2013) used this technique to distinguish how technology could be integrated with teaching strategies to develop medical proficient practitioners. With the intention of teachers' competencies, it has been used to identify, develop, and validate competences framework of teaching in different subjects namely English, physical education, mathematics, science, and counselor educators (Alaa et al., 2019; Afandi et al., 2019; Muñiz-Rodríguez et al., 2017; Swank & Houseknecht, 2019; Wyant et al., 2020). Accordingly, the current study adopted this technique to find out a consensus among some experts in several fields about the required competencies of professional faculty in Arab universities.

## **2. Background**

The urgency of accelerating the digital transformation of education requires a paradigm shift in how we understand education and learning. Faced with the pandemic caused by Covid-19, online education is presented as a necessary response and, in order to successfully enter it, we compile the keys to this modality according to our experience as an educational institution of online teaching. Many challenges have been represented that may saddle faculty in higher education institutions. They need to keep pace with the innovative paradigms of higher education, new approaches to teaching and learning, and how the online tools can be used to support the instruction activities (Siemens & Matheos, 2010; Albrahim, 2020). Universities should invest in teacher professional development of their faculty, now more than ever, for them to be updated on effective pedagogical methods with or without the use of online technologies (Rapanta et al., 2020). Robinson (2017) pointed out, in his research on examining the quality measurement standards by online instructors, that a disparity between the expectations of the creation, development, and application of online courses that are not typically expected of onsite courses.

From its birth in the last decades to the present day, online courses and education platforms have been an open option for millions of students around the world; However, the situation of the pandemic we are going through as a society, which brought at least half of the students and professionals of education all over the planet home in a matter of four months, has once again raised the digitization of education not as one more option, but as a necessity both for educational institutions, companies and students (Yousef & Sumner, 2020). Further, it must be recognized that the socio-economic and socio-educational realities are not the same in all cultures. Therefore, each institution must design, as far as possible, online teaching models according to the socio-educational and socio-economic indicators of the community (Kem-mekah Kadzue, 2020). Faculty need also to keep themselves updated with the dynamic nature of online learning and emerging learning technologies and mode of teaching and learning in virtual environments (Kibaru, 2018). As such, it is important for faculty to perceive and use technology as an integral part of a student-centered approach to teaching if enhanced learning outcomes are to be achieved (Englund et al., 2017; Kreber & Kanula, 2013).

The existing literature base contains several studies on measuring the quality of teaching and learning courses by adopting some standards of course design, curriculum, and assessment tools. Of the studies reviewed that focused specifically on professional development through online teaching, Frankel (2015) addressed high-quality professional development and mentoring activity for online is essential to educational systems, it needs to be supplemented by intuitive feedback that leads to a planned set of professional learning activities to help faculty improve their practice. An earlier study investigated the value of contextualization, incremental innovation, and mentoring of online convenors. it concluded that teaching online or blended types of learning needs to be rapid, cost-effective, and lead directly to practical outcomes (Gregory & Salmon, 2013). The findings of another recent research, that investigated the design of online learning activities by using certain features, concluded the need for adjusting assessment to the new learning needs, and the sequence of three types of faculty presence namely social, cognitive, and facilitator (Rapanta et al., 2020).

In their report, Ni She et al. (2019) emphasized three key elements for effective teaching online namely presence, facilitation, and supporting students. These elements have mapped 18 associated core competencies of online educators in seven main roles managerial, pedagogical, social, technical, assessor, facilitator, and content expert. Closely related to investigating the Arab higher education institutions, a recent study concluded six main skills that faculty members need to efficiently teach in online learning environments which have to be determined in order to help design professional development programs for online instructors (Albrahim, 2020). Further, another study referred that online classrooms may seem inherently anti-social, leaving many faculty members wondering how to best approach discussion between students, and how to use collaboration for more work-intensive tasks. Moreover, faculty members must also develop alternate strategies to make sure students are progressing in the course (Abell et al., 2016).

Williams (2003) concluded thirteen distinct roles are needed to implement distance education programs in higher education (e.g., administrative manager, technology expert, librarian, evaluation specialist, and leader/change agent), the author highlighted the importance of interpersonal-related and communication-related skills between university teachers and their students in this type of educational environment. Five main roles, which could be identified with regards to the tasks carried out by university teachers in online environments, were reported by Alvarez et al. (2009) namely designer or planning, social, cognitive, technological, and managerial roles.

Some previous studies have argued the required competencies or skills for some specific programs and courses. O'Doherty et al. (2019) argued the internet skills of faculty in medical fields during the implementation of online and distance learning methodologies. they concluded some requirements of specific creative skills, information navigation, and social media training, in order to address many of the challenges faced in an expanding digital world. An investigation of teaching with technology in a Master programme of Pharmacy at a Swedish university indicated clear differences between novice and experienced teachers. the novice teachers demonstrated greater and more rapid change in practices of teaching with technology than experienced colleagues. the experienced teachers tended to exhibit little to no change in conceptions (Englund et al., 2017). A recent study explored the perceived roles and competencies of e-tutors in the economic and management sciences college, the findings concluded that faculty perceived a challenge to engage learners in online settings, and it highlighted the importance of the social or pastoral roles of the faculty on the successful online interactions with students (De Metz & Bezuidenhout, 2018).

Indeed, online teaching is not the same as face-to-face classes, they are different formats, and processes that have different logics and structures. The core challenge is that the face-to-face educator believes that doing virtual education is simply transferring the same concepts, structures, and class organization from the face-to-face space to virtual space, and this is not the case. It is a new design, logic, and structure (Ni She et al., 2019; Kibaru,

2018; Trammell & LaForge, 2017). In exploring the issues, needs, and outcomes of government organizations in developing countries seeking to implement information technology into teaching and learning practices, Passey et al. (2016) concluded it should be recognized that patterns of support, for those working in these countries, have not been achieved at any identifiable widespread level. In the case of many Arab universities, faculty were forced to teach online, and many faculties lack some of the requirements and competencies to teach professionally online. Albrahim (2020) concluded, in his investigation on Arab higher education institutions, that Arab faculty members might feel uncomfortable and not familiar with online teaching courses due to the multiple roles and responsibilities of teaching online. In this context, while online learning may lend itself to independent student learning, some students need hands-on, interactive tasks to engage and challenge them (Abell et al., 2016). They struggle also with many obstacles relevant to the technical infrastructure. Therefore, one of the main challenges is to achieve an efficient course design, making the most of the tools that this modality allows, adequately planning the contents, evaluation activities, and student dedication times, as well as the continuous support and monitoring of the Teacher. Taking into consideration the number of students in most Arab universities, course design and instructional effectiveness are some of the most significant challenges facing faculty tasked with managing large online courses (Trammell & LaForge, 2017). Hence, it will increase extra challenges particularly with the common learning styles of Arab students. Ultimately, faculty need some personal and professional attributes, skills, and knowledge which can be the core competencies for them in supporting and developing their professional roles.

### 3. Method

#### 3.1. Study design

The Delphi technique is an iterative process for analyzing the opinions of many experts based on the outcomes of several questionnaire rounds (Saffie & Rasmani, 2016). For predictable content analysis, this study used a Delphi technique in a series of rounds or sequential questionnaires that were hosted by QuestionPro, intermixed by structured feedback after each round as depicted in Figure 1.

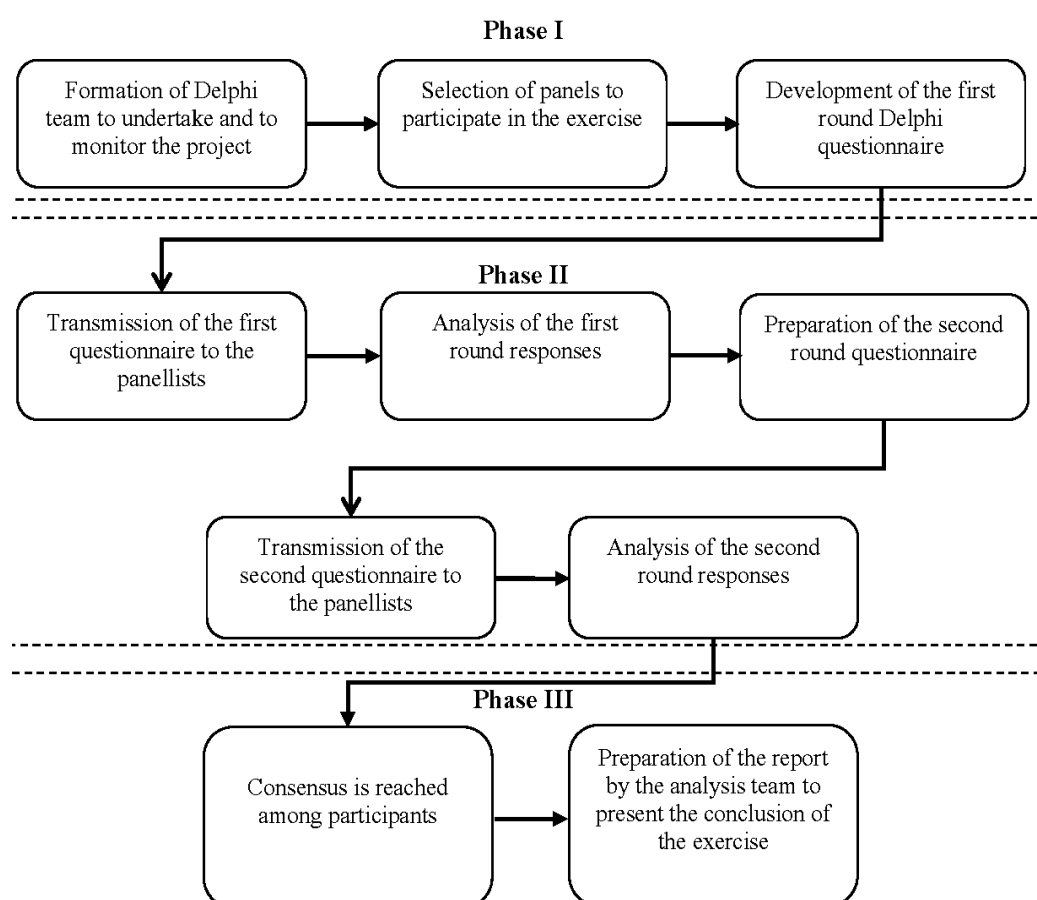


Figure 1. Delphi technique in a series of rounds (Saffie & Rasmani, 2016)



The data collection began with a short introduction to the study included a brief description of three levels of required competencies for the professionalism of online teaching namely, knowledge, skills, and attributes. Theoretically, the Delphi technique could highlight the areas of divergence of opinions, so the combined opinions of experts are of richer quality than the limited view of an individual (Nworie, 2011). Consequently, the current study used this technique to evoke experts' perspectives on how faculty can be professional online teachers through their attributes, challenges, and training needs. Considering this, three rounds were planned to align with the components of the Delphi technique by answering the following questions:

- What are the competencies for the professionalism of online teaching? (Round 1)
- What are the expected roles of faculty for online teaching? (Round 2)
- What are the knowledge, skills, and attributes for each role? (Round 3)

Data collected establishes consensus among experts in an iterative aspect, the stimulus of each round was decided based on anonymous responses of the previous round. The storyboard of three rounds can be summarized as the following:

- **Round 1:** Establishing a list of expected knowledge, skills, and attributes that experts consider necessary for professional online teaching. Three open-ended questions were given to panelists regarding: What are the knowledge, skills, and attributes that faculty members need to be professional for online learning? This round intended to gather all possible requirements of online teaching professionalism through different roles before, during, after sessions.
- **Round 2:** The second phase was designed based on the synthesis of ideas that developed from Round 1. The experts had the opportunity to classify the listed knowledge, skills, and attributes from Round 1 into three main roles namely as a planner, assessor, and mentor. The created list for each role, by quantitative analysis of responses, was carried forward to be more inspected and investigated in the final round.
- **Round 3:** In this stage, the panelists were given an occasion to reconsider their answers by ranking the top 10 items, in order of the importance of each role, to create a matrix of 10 required knowledge, skills, and attributes of each role.

### 3.2. Participants

A total number of 29 experts in different disciplines were asked to participate in the first round. Table 1 shows the response rate of each round and the range of their disciplines and academic experience of teaching in higher education institutions. There was no direct communication between experts on the study subject and none of them was aware of the list of participants.

*Table 1. Characteristics of the panelist participated in the study rounds*

Response rate and characteristic	No. of respondents (%)
Response rate	
Round 1	24/29 (82.7%)
Round 2	21/24 (87.5%)
Round 3	21/21 (100 %)
Years of academic experience (Overall $N = 24$ )	
≤ 10	34.4%
> 10	65.6%
Academic disciplines (Overall $N = 24$ )	
Huminites and Educational sciences	41.5%
Basic Sciences	25%
Computer Sciences	21 %
Medical Sciences	12.5 %

### 3.3. Data analysis

The study used different data analysis tools for each round. In the first round, to list the requirements of effective online teaching, a qualitative content analysis of answers was carried out independently by each author. Each author analyzed the responses in order to list the potential knowledge, skills, and attributes according to the panelists. A third reviewer was also consulted to eradicate any bias of authors in analyzing the responses. The

statistical analyses were used in the second round. The median and frequency distribution values, by using the fifth Likert scale, were calculated to inspect the level of agreement on items for each role of faculty during online teaching namely as a planner, an assessor, and a mentor. The final round aimed to rank the top 10 required competencies for each role. The points for each item were allocated by the total of all experts as follows: 10 points for the first order, 9 points for the second order, continuing to the last order by 1 point, and the value of 0 was given if an item did not occur in the top 10. This method to achieve consensus among experts was used by many researchers (Moynihan et al., 2015; Milat et al., 2013).

## 4. Findings

**Round 1:** The experts reported 59 requirements for professional online teaching. A list of 18 knowledge, 22 skills, and 19 attributes was classified as considered necessary items for the efficiency of faculty members in online learning environments. The reported list of knowledge included the staff ability to recognize the core outcomes of sessions, the available applications or platforms for assessment tasks, the quick solutions and IT support services, and the extra resources for students. Moreover, outline the mechanisms of formative evaluation, engaging and motivating environment for students, team building, and statistical analysis of the platforms. At the level of skills, 22 skills were listed to reflect the professional practices of online teaching.

**Round 2:** Constructing on the insights from the first round, 21 experts participated in the second round to create the required knowledge, skills, attributes for online teaching professionalism to measure the role of faculty as a planner, assessor, and mentor. Table 2 shows the median of each item and frequency distribution of knowledge, skills, attributes for each role. Seven items achieved a median of <5 and a frequency distribution of <50% and were dismissed in the third round. Consequently, the analysis of this stage formed 13 knowledge, 12 skills, and 14 attributes issued to participants in the third round.

*Table 2. List of reported knowledge, skills, and attributes from Round 2*

Items (*)	Median	Frequency	CD
<b>As Planner</b>			
Effectively use crisis management techniques.	9.2	95.2%	<b>A</b>
Identify the advantages and disadvantages of electronic platforms/ applications.	9.2	90.5%	<b>K</b>
Use communication skills effectively by different channels/medium.	8.9	90.5%	<b>S</b>
Managing time effectively during the sessions.	8.6	90.5%	<b>S</b>
Develop the personal ability to use modeling in educational situations.	8.6	90.5%	<b>A</b>
Use the appropriate applications or programs for subjects and educational goals.	9	85.7%	<b>A</b>
Effectively use flipped classroom techniques to plan course sessions.	8.3	81%	<b>A</b>
Design a plan for each lesson by using different strategies / electronic tools.	7.7	81%	<b>S</b>
Effectively use the tone of voice during the session.	7	81%	<b>S</b>
Identify the appropriate applications consistent with the limitations speed of the available internet connection /network.	6.4	71.4%	<b>K</b>
Effectively use, in case of the technical problems, the alternatives to learning activities.	6.1	66.6%	<b>A</b>
Establish an activity bank of lessons by using various electronic tools.	5.5	61.9%	<b>S</b>
Recognize the quick solutions/instructions for technical problems.	5.7	57.1%	<b>K</b>
Recognize the core outcomes at the level of course and sessions.	5.3	52.4%	<b>K</b>
Use attractive videos and pictures during the session. (**)	4.4	42.9%	<b>S</b>
Being team-oriented in teaching style. (**)	3.8	42.9%	<b>A</b>
<b>As Assessor</b>			
Identify effective tools to create a more engaging and motivating environment for students.	9.1	95.2%	<b>K</b>
Use different monitoring techniques to measure students' contributions during sessions.	9.7	95.2%	<b>S</b>
Outline the mechanisms of formative evaluation which can be used during online sessions.	8.6	90.5%	<b>K</b>
Use effective techniques, during sessions, to measure students' attention.	8.6	90.5%	<b>S</b>
Demonstrate effective tools to motivate his inactive students during the sessions.	8.6	90.5%	<b>S</b>
Provide constructive and continuous feedback on students' interactions during the session.	8.8	90.5%	<b>A</b>
Use effective reflections on students' performances and achievements.	8.6	85.7%	<b>A</b>

Identify students' limitations of using the available applications/tools for evaluation and assessment.	7.4	76.2%	K
Effectively use flipped classroom techniques to measure students' performance	7.4	76.2%	A
Outline the appropriate tools to assess each outcome/objective.	6.7	66.6%	K
Use the tools of alternative assessment for different learning outcomes.	6.6	66.6%	S
Use creative alternatives to increase students' performance/achievement.	5.6	57.1%	A
Effectively use activities to measure the students' achievement of outcomes.	5.1	52.4%	A
Recognize the available applications for evaluating students' achievement. (**)	4.8	47.6%	K
Establish a question bank to measure lessons learning outcomes during the sessions. (**)	4.6	47.6%	S
Recognize the available statistical analysis of the platforms/ applications. (**)	3.7	42.9%	K
<b>As Mentor</b>			
Create a constructive environment in which motivate students to participate effectively.	9.1	95.2%	A
Encourage the students to share their needs and academic obstacles.	9.1	90.5%	S
Apply effective tools of class management.	8.7	90.5%	S
Lead effectivity team-work discussions	8.9	90.5%	A
Identify students learning styles and preferences.	8.4	85.7%	K
Recognize the available IT support services for students	9.1	85.7%	K
Show academic commitment towards learners needs	8.8	85.7%	A
Identify the extra resources needed / available for students.	7.2	76.2%	K
Demonstrate effective strategies for problem-solving in instructional tasks.	7.9	76.2%	S
Identify the academic history of students and their previous achievements.	7.1	66.6%	K
Identify team building techniques/tools.	5.3	52.4%	K
Use students' feedback to develop an action plan for the effectiveness of teaching.	5.4	52.4%	A
Seeing students/other points of view. (**)	3.8	38.1%	S
Show a positive relationship with students. (**)	4.1	38.1%	A

Note. (\*) Items have been ordered by frequency distribution per each role. (\*\*) Was Eliminated from Round 3. CD = Category Domain; K= Knowledge; S= Skills; A= Attributes.

- **Round 3:** In the final round, 21 experts were asked to rank the revised list of knowledge, skills, attributes from Round 2. A matrix of 30 requirements, as shown in Table 3 below, is outlined of the top required knowledge, skills, attributes for faculty as a planner, assessor, and mentor in online teaching.

Table 3. Matrix of 30 requirements for online teaching professionalism

	Planner	Assessor	Mentor
Knowledge	<ul style="list-style-type: none"> <li>• Identify the advantages and disadvantages of electronic platforms/ applications.</li> <li>• Identify the appropriate applications consistent with the limitations speed of the available internet connection /network.</li> <li>• Recognize the core outcomes at the level of course and sessions.</li> </ul>	<ul style="list-style-type: none"> <li>• Identify effective tools to create a more engaging and motivating environment for students.</li> <li>• Outline the mechanisms of formative evaluation which can be used during online sessions.</li> <li>• Identify students' limitations of using the available applications/tools for evaluation and assessment.</li> </ul>	<ul style="list-style-type: none"> <li>• Identify students learning styles and preferences.</li> <li>• Identify the extra resources needed / available for students.</li> <li>• Recognize the available IT support services for students.</li> <li>• Identify the academic history of students and their previous achievements.</li> </ul>

Skills	<ul style="list-style-type: none"> <li>• Use communication skills effectively by different channels/medium.</li> <li>• Managing time effectively during the sessions.</li> <li>• Design a plan for each lesson by using different strategies / electronic tools.</li> <li>• Effectively use the tone of voice during the session.</li> </ul>	<ul style="list-style-type: none"> <li>• Use different monitoring techniques to measure students' contributions during sessions.</li> <li>• Use effective techniques, during sessions, to measure students' attention.</li> <li>• Demonstrate effective tools to motivate his inactive students during the sessions.</li> </ul>	<ul style="list-style-type: none"> <li>• Encourage the students to share their needs and academic obstacles.</li> <li>• Apply effective tools of class management.</li> <li>• Demonstrate effective strategies for problem-solving, completion of educational tasks.</li> </ul>
Attributes	<ul style="list-style-type: none"> <li>• Effectively use crisis management techniques.</li> <li>• Use the appropriate applications or programs for subjects and educational goals.</li> <li>• Effectively use flipped classroom techniques to plan course sessions.</li> </ul>	<ul style="list-style-type: none"> <li>• Provide constructive and continuous feedback on students' interactions during the session.</li> <li>• Use effective reflections on students' performances and achievements.</li> <li>• Effectively use flipped classroom techniques to measure students' performance.</li> <li>• Effectively use activities to measure the students' achievement of outcomes.</li> </ul>	<ul style="list-style-type: none"> <li>• Create a constructive environment in which motivate students to participate effectively.</li> <li>• Lead effectivity team-work discussions.</li> <li>• Show academic commitment to learners' needs.</li> </ul>

## 5. Discussion

The consensus among experts reflects three levels of required competencies for the staff to be a professional teacher. The listed knowledge and skills, as reported in the first round, included the minimum level of dealing with the expected technical problems, and the advantages and disadvantages of electronic applications. Moreover, their ability to use effective communication skills, managing time, monitoring and motivating techniques, class management, and problem-solving skills. The experts also reported some of the skills relevant to establish a question bank, alternative assessment tools, activity bank, and lesson plan by using different electronic tools. For the professional attributes, the panelist reported some attributes to measure the ability of staff to provide constructive feedback, effective reflections on students' performance, and an attractive online environment. Some techniques were listed relevant to use modeling in educational situations, flipped classroom techniques, team-oriented of teaching style, and crisis management skills.

The revised list referred to some of the required knowledge, skills, and attributes for the professionalism of online teaching in each role of the educational processes. As a planner, the results listed the essential ability to use crisis management techniques, create an effective environment of learning by recognizing the advantages and disadvantages of the available tools, and choose the appropriate applications based on the limitations speed of the available internet connection. Moreover, identify students' learning styles and their preferences in order to use the effective tools and mechanisms for motivating students, and recognize the formative evaluation tools and flipped classroom techniques that can be used during sessions. For the role as an assessor, the panelist reported that the faculty should find the effective tools of engaging, motivating, measuring students' attention, and monitoring their progress. Additionally, provide their reflections on students' performances during the session and use effectively the tools of alternative assessment for different learning outcomes. As mentors, the findings exposed the relative importance of the faculty role in encouraging the students to share their needs and academic obstacles, applying the effective tools of class management and problem-solving, and showing the academic commitment towards learners' needs.

The requirements of professional online teaching were concluded by 30 practices in three expected roles of faculty. The findings pointed out that professional online teaching comprises a level of knowledge to recognize the properties of the available electronic platforms or applications consistent with its advantages or disadvantages, and the limitations of using the internet. Furthermore, faculty should be aware of some characteristics of their students e.g., their learning styles and preferences, academic history, and previous achievements. The experts reported also required knowledge of the mechanisms of formative evaluation, the available IT support services, and the extra resources which can be used to measure the core outcomes of the course. Several skills have been reported for determining the ability of faculty to use effective communication

skills, monitoring and motivation techniques, class management, problem-solving skills, and time management skills.

A list of 10 important features was cataloged as attributes of professional online faculty to measure their competencies of using crisis management techniques, using flipped classroom, providing constructive feedback and motivating environment, choosing the appropriate applications, reflecting on students' performances and achievements, leading team-work discussions. By considering the discipline-specific competencies and investigating the listed competencies according to the academic disciplines of experts, it is clear to find a consensus among experts in the field of humanities and education sciences about competencies related to class management skills and how teachers can motivate their students in the conditions of online classes. For basic sciences experts, they focused on monitoring techniques and effectively planning of sessions' activities. Rationally, the problem-solving skills and using the appropriate applications on educational activities were adopted by the experts in computer sciences field who determined teacher digital competencies (TDC) to recognize the progressively complex knowledge, skills, and attributes of teachers to deal with students' needs of learning ethically, safely, and productively in a varied digital environment. The reported TDC aligned with Falloon's conclusions, in his framework of the successful teachers' transition from digital literacy to digital competence, that highlighted the importance of several competencies in classroom roles of teachers through modeling and deliberate planning to educate their students in building the ability to leverage advantage from digital resources by sustainable ways (Falloon, 2020).

These findings are partly mirrored with the conclusions found in a previous study by Lee and Hirumi (2004) which presented sixteen outputs for performing six main skills namely interaction, management organization/instructional design, technology, content knowledge, and teamwork skills. The study findings also somewhat consistent with previous studies which concluded some categories of skills and competencies required for teaching online courses in higher education e.g., pedagogical skills, content skills, monitor students' progress, design skills, responsiveness, technological skills, encourage active learning, management and institutional skills, and social-communication skills (Albrahim, 2020; Ni She et al., 2019; Alvarez et al., 2009). Besides, the reported attributes in the existing study align with the conclusion of Rose's study that effective online teachers need to avoid a didactic approach, which is lecture-based, and provide a seamless structure by using actively engage students, establishing a learning-oriented social presence online (Rose, 2018). Bigatel et al. (2012) also reported some of the listed skills and knowledge of the present study in their identification of competencies for online teachers as "connectors" between the learner and his or her learning system by labeling active learning to construct explanations, solutions, hands-on practice, student-generated content.

Eventually, online teaching competencies are satisfactorily documented across the literature where certain teachers' skills and attributes have been researched, but the existing study distinguishes between three different levels of professional competencies. Moreover, the study investigated how the required competency can be changed based on three main roles or responsibilities of professional online teachers. In addition, the study strengthens the idea that online learning activities can be used to enhance teaching and knowledge sharing between teachers and students. There are two types of effect which result when students utilize these learning activities. Firstly, teachers can involve students in the instructional process using relevant activities and discussions from any convenient place at any time. Secondly, it promotes additional learning experiences where students can interact, collaborate, and take ownership of their learning. Where students can share ideas, experiences, perspectives, and opinions that support self-self-directed and collaborative content sharing.

## 6. Conclusion

The study aimed to determine the capability and suitability of faculty members at universities to be a professional online teacher. The range of requirements, that were assumed most important by experts, reflects that online teaching professionalism is not a simple set of conduct, behavior, or attitude. It encompasses several different attributes that define the professional skills and a minimum requirement of knowledge. The final list of requirements, that were deemed to have achieved consensus as to its importance, shows the readiness for online teaching by six main domains namely evaluating students' achievements and limitations, problem-solving skills, IT, and computer skills, monitoring and motivating techniques, communication skills, and class management skills.

The suggested matrix can be used to measure the professionalism of online teaching in three main roles of faculty namely as a planner of learning activities or scenarios, as an assessor of students' achievements and progress, and as a mentor of coaching and motivating activities. The results can be used as a part of professional



development programs to provide faculty with the skills to meet the standards of online teaching professionally. The findings also recognized faculty attributes through strengthening perceptible skills and recognizing less palpable personal characteristics, which can perhaps contribute more meaningfully to enhancing the efficacy of online teaching.

The scope of this study was limited in terms of the study sample and culture. The study sample did not include all disciplines, but it was limited to some specialties in different disciplines i.e., humanities and educational sciences, basic sciences, computer sciences, and medical sciences. Moreover, this research was conducted in the Arab culture and the identifications made by the authors could be perceived differently in other cultures. In terms of the study design, the authors used Delphi technique, accordingly, future studies may consider other techniques to compare the findings. Different levels of competencies and roles can be also investigated as open points. Consequently, the required competencies might change under different conditions e.g., students' background, heterogeneous groups, ICT knowledge or skills, and appropriate social media. Thus, the upcoming studies can provide more nuanced and direct evidence of whether faculty adopted changes in perspectives, principles, and intentions for developing in their performs were realized in online teaching.

## References

- Abell, N., Cain, M., & Lee, CYC. (2016). Essential attributes for online success: Student learning preferences and faculty teaching styles. *International Journal on E-Learning*, 15(4), 401-422.
- Afandi, A., Sajidan, S., Akhyar, M., & Suryani, N. (2019). Development frameworks of the Indonesian partnership 21st-century skills standards for prospective science teachers: A Delphi study. *Journal Pendidikan IPA Indonesia*, 8(1), 89-100. doi:10.15294/jpii.v8i1.11647
- Alaa, M., Albakri, Intan Safinas Mohd Ariff, Singh, C. K. S., Hamed, H., Zaidan, A. A., Zaidan, B. B., Albahri, O. S., Alsalem, M. A., Salih, M. M., Almahdi, E. M., Baqer, M. J., Jalood, N. S., Nidhal, S., Shareef, A. H., & Jasim, A. N. (2019). Assessment and ranking framework for the English skills of pre-service teachers based on fuzzy Delphi and TOPSIS methods. *IEEE Access*, 7, 126201-126223. doi:10.1109/ACCESS.2019.2936898
- Albrahim, F. A. (2020). Online teaching skills and competencies. *TOJET the Turkish Online Journal of Educational Technology*, 19(1), 9-20.
- Alvarez, I., Guasch, T., & Espasa, A. (2009). University teacher roles and competencies in online learning environments: A Theoretical analysis of teaching and learning practices. *European Journal of Teacher Education*, 32(3), 321-336. doi:10.1080/02619760802624104
- Badia, A., Garcia, C., & Meneses, J. (2017). Approaches to teaching online: Exploring factors influencing teachers in a fully online university: Factors influencing approaches to teaching online. *British Journal of Educational Technology*, 48(6), 1193-1207. doi:10.1111/bjet.12475
- Bigatel, P. M., Ragan, L. C., Kennan, S., May, J., & Redmond, B. F. (2012). The Identification of competencies for online teaching success. *Journal of Asynchronous Learning Networks JALN*, 16(1), 59-77.
- De Metz, N., & Bezuidenhout, A. (2018). An Importance-competence analysis of the roles and competencies of e-tutors at an open distance learning institution. *Australasian Journal of Educational Technology*, 34(5). doi:10.14742/ajet.3364
- Englund, C., Olofsson, A. D., & Price, L. (2017). Teaching with technology in higher education: Understanding conceptual change and development in practice. *Higher Education Research and Development*, 36(1), 73-87. doi:10.1080/07294360.2016.1171300
- Falloon, G. (2020). From digital literacy to digital competence: The Teacher digital competency (TDC) framework. *Educational Technology Research and Development*, 68(5), 2449-2472. doi:10.1007/s11423-020-09767-4
- Frankel, C. E. (2015). *Online teaching: Professional development for online faculty*. ProQuest Dissertations Publishing (UMI: 3682679).
- Green, R. A. (2014). The Delphi technique in educational research. *SAGE Open*, 4(2), 215824401452977. doi:10.1177/2158244014529773
- Gregory, J., & Salmon, G. (2013). Professional development for online university teaching. *Distance Education*, 34(3), 256-270. doi:10.1080/01587919.2013.835771
- Kem-mekah Kadzue, O. (2020). Online teaching during the Covid-19 crisis in Cameoon's University Education: Achievements and challenges. *EHQUIDAD. International Journal of Welfare and Social Work Policies*, (14), 57-74. doi:10.15257/ehquidad.2020.0012

- Kibaru, F. (2018). Supporting faculty to face challenges in design and delivery of quality courses in virtual learning environments. *The Turkish Online Journal of Distance Education TOJDE*, 19(4), 176-197. doi:10.17718/tojde.471915
- Kreber, C., & Kanuka, H. (2013). The Scholarship of teaching and learning and the online classroom. *Canadian Journal of University Continuing Education*, 32(2), 109-131. doi:10.21225/D5P30B
- Lee, J.-L., & Hirumi, A. (2004, October). *Analysis of essential skills and knowledge for teaching online*. Chicago, IL: Paper Presented at the Association for Educational Communications & Technology (ED485021).
- Milat, A. J., King, L., Bauman, A. E., & Redman, S. (2013). The Concept of scalability: Increasing the scale and potential adoption of health promotion interventions into policy and practice. *Health Promotion International*, 28(3), 285-298. doi:10.1093/heapro/dar097
- Mirata, V., Hirt, F., Bergamin, P., & van der Westhuizen, C. (2020). Challenges and contexts in establishing adaptive learning in higher education: Findings from a Delphi study. *International Journal of Educational Technology in Higher Education*, 17(1), 1-25. doi:10.1186/s41239-020-00209-y
- Mishra, L., Gupta, T., & Shree, A. (2020). Online teaching-learning in higher education during lockdown period of COVID-19 pandemic. *International Journal of Educational Research Open*, 1. doi:10.1016/j.ijedro.2020.100012
- Moynihan, S., Paakkari, L., Välimaa, R., Jourdan, D., & Mannix-McNamara, P. (2015). Teacher competencies in health education: Results of a Delphi study. *PloS One*, 10(12), e0143703-17. doi:10.1371/journal.pone.0143703
- Muñiz-Rodríguez, L., Alonso, P., Rodríguez-Muñiz, L. J., & Valcke, M. (2017). Developing and validating a competence framework for secondary mathematics student teachers through a Delphi method. *Journal of Education for Teaching: JET*, 43(4), 383-399. doi:10.1080/02607476.2017.1296539
- Ni She, C., Farrell, O., Brunton, J., Costello, E., Donlon, E., Trevaskis, S., Eccles, S. (2019). *Teaching online is different: critical perspectives from the literature*. Dublin, Ireland: Dublin City University. doi:10.5281/zenodo.3479402
- Nworie, J. (2011). Using the Delphi technique in educational technology research. *Techtrends*, 55(5), 24-30. doi:10.1007/s11528-011-0524-6
- O'Doherty, D., Loughheed, J., Hannigan, A., Last, J., Dromey, M., O'Tuathaigh, C., & McGrath, D. (2019). Internet skills of medical faculty and students: Is there a difference? *BMC Medical Education*, 19(1), 39. doi:10.1186/s12909-019-1475-4
- Passey, D., Laferrière, T., Ahmad, M. Y. A., Bhowmik, M., Gross, D., Price, J., Resta, P., & Shonfeld, M. (2016). Educational digital technologies in developing countries challenge third party providers. *Educational Technology & Society*, 19(3), 121-133.
- 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. doi:10.1007/s42438-020-00155-y
- Robinson, D. J. (2017). *A Delphi study to examine the quality measurement standards by online instructors using the quality matters™ rubric as a basis for creating instructional materials* (Unpublished doctoral dissertation). Kent State University, Kent, OH.
- Rose, M. (2018). What are some key attributes of effective online teachers? *Journal of Open, Flexible, and Distance Learning*, 22(2), 32-48.
- Rowe, M., Frantz, J., & Bozalek, V. (2013). Beyond knowledge and skills: The Use of a Delphi study to develop a technology-mediated teaching strategy. *BMC Medical Education*, 13(1), 51. doi:10.1186/1472-6920-13-51
- Saffie, N. A. M., & Rasmani, K. A. (2016, July). Fuzzy Delphi method: Issues and challenges. In *2016 International Conference on Logistics, Informatics and Service Sciences (LISS)* (pp. 1-7). doi:10.1109/LISS.2016.7854490
- Shelly, J. L., & Scolaro, K. L. (2016). Utility of a professionalism assessment form activity: A Survey of students and teaching assistants. *Currents in Pharmacy Teaching and Learning*, 8(1), 119-124. doi:10.1016/j.cptl.2015.09.015
- Siemens, G., & Matheos, K. (2010). Systemic changes in higher education. In *Education*, 16(1), 3-18. Retrieved from <http://ineducation.ca/ineducation/article/view/42>
- Sims, T. N. (2017). *A Qualitative study to understand high school teachers' experiences teaching online* (Unpublished doctoral dissertation). Capella University, Minneapolis, MN.
- Swank, J. M., & Houseknecht, A. (2019). Teaching competencies in counselor education: A Delphi study. *Counselor Education and Supervision*, 58(3), 162-176. doi:10.1002/ceas.12148
- Trammell, B. A., & LaForge, C. (2017). Common challenges for instructors in large online courses: Strategies to mitigate student and instructor frustration. *The Journal of Educators Online*, 14(1).
- United Nations Educational, Scientific and Cultural Organization (UNESCO). (2020). *290 million students out of school due to COVID-19: UNESCO releases first global numbers and mobilizes response*. Retrieved from

<https://en.unesco.org/news/290-million-students-out-school-due-covid-19-unesco-releases-first-global-numbers-and-mobilizes>

Williams, P. E. (2003). Roles and competencies for distance education programs in higher education institutions. *The American Journal of Distance Education*, 17(1), 45-57. doi:10.1207/s15389286ajde1701\_4

Wyant, J. D., Tsuda, E., & Yeats, J. T. (2020). Delphi investigation of strategies to develop cultural competence in physical education teacher education. *Physical Education and Sport Pedagogy*, 25(5), 525-538. doi:10.1080/17408989.2020.1746252

Yousef, A. M. F., & Sumner, T. (2020). Reflections on the last decade of MOOC research. *Computer Applications in Engineering Education*. doi:10.1002/cae.22334

## Exploring Effects of Geometry Learning in Authentic Contexts Using Ubiquitous Geometry App

Wu-Yuin Hwang<sup>1</sup>, Uun Hariyanti<sup>1\*</sup>, Yan Amal Abdillah<sup>1</sup> and Holly S. L. Chen<sup>2</sup>

<sup>1</sup>Graduate Institute of Network Learning Technology, National Central University, Jhongli City, Taiwan //

<sup>2</sup>Luzhou Elementary School, Taipei City, Taiwan // wyhwang@cc.ncu.edu.tw //

uun.hariyanti1802@g.ncu.edu.tw // yan.amal91@gmail.com // hollyc4826@gmail.com

\*Corresponding author

(Submitted May 20, 2020; Revised July 4, 2020; Accepted February 25, 2021)

**ABSTRACT:** Geometry is essential for mathematics learning given that it is strongly related to our surroundings; however, few studies concentrated on using geometry in our daily life, especially using mobile devices with their sensors. Thus, this study proposed one app, Ubiquitous Geometry (UG), and explored its effects on learning angles and polygons in authentic contexts. The experiment was conducted for grade four learners of an elementary school. The control group used protractors and pencil/paper in measuring angles and polygons, whereas the experimental group did measurements with UG. The results showed that in terms of learning achievement, the experimental group outperformed the control group. Further investigation of the relationship between learning behaviors and learning achievement in the experimental group found that both learning effectiveness and quantity of learning, including measuring angles of elevation and depression (MED), note drawing, and comment drawing, have significantly positive correlations with learning achievement. These three behaviors also become significant predictors of learning achievement after multiple regression analysis. Moreover, MED was found to be the most critical factor to affect learning achievement. Additionally, in perception evaluation, participants felt satisfied with UG and authentic measurement activities by which their learning motivation and interests in authentic contexts were indeed stimulated. Hence, we suggested that UG was worth promoted and further investigated its effects on authentic geometry learning.

**Keywords:** Measurement in authentic contexts, Learning behaviors, Cognitive abilities, Ubiquitous Geometry

### 1. Introduction

In terms of learning, authentic contexts are not as simple as applying real-life practices. The contexts should be based on learning purpose, motivation, and complex learning environment so that these can be explored and applied by learners in their surroundings (Herrington & Kervin, 2007). This can be done by providing learners with real-life problems that can be explored to promote a better learning process. Hence, each experience in the learning process, especially in math, should be aimed to inculcate real-life applications into every task, lesson, and unit to enhance the cognitive development and master learning abilities through failure experiences and more practices (Nicaise, Gibney, & Crane, 2000).

Two educational theories related to the learning process are taken to underpin this study. The first is enactivism, a combination of constructivism and embodied cognition, which holds cognition and environment to be inseparable (Ernest, 2010). Learning, then, occurs when learners interact with their environments. By following this theory, applying real-life problems into learning tasks will make learners do authentic activities and interact with environments. The authentic activity is used to encourage learners' participation (Herrington, Oliver, & Reeves 2003) to apply their knowledge in their surroundings (Hwang et al., 2011; Hwang et al., 2015; Hwang et al., 2019). It can be used as an essential factor in assessing mathematics learning behaviors (Wang et al., 2016) and enhancing cognitive levels (Kong, Wong, & Lam, 2003). Thus, we focused on applying authentic activities in learning geometry, mainly for measuring and learning angles and polygons in surroundings; hopefully, it can enhance students' learning behaviors and help their learning performance as well.

The second theory is social constructivism, which is a combination of the idea of social interaction (on the social level) and learning by doing (on the individual level) to make learning more meaningful and enhance cognitive development (Vygotsky, 1978). Social interaction has a mediation role in which learners could perform successful tasks when they are in the interaction of giving or receiving help to or from other learners. Several studies applied social interaction by applying peer assessment in classroom practice (Barak & Asakle, 2018; Lai & Hwang, 2015). Peer assessment and peer comments, which are a part of social interaction, can improve learning interaction and provide in-depth knowledge by making reflections and doing communication with others (Chung, Hwang, & Lai, 2019; Engeström, 1999). The learning improvement can be satisfied whereby the

peer can give helpful comments related to the problems or contexts (Hwang & Hu, 2013). Therefore, as claimed by Vygotsky's (1978), social interaction in peer assessment and peer comments between two or more learners with different levels of skills and knowledge is the core attribute of effective learning. Moreover, doing authentic activities (e.g., measuring real objects), which is an implication of learning by doing, can increase learners' engagement and value of comments. By reflecting on knowledge received in real objects measurements and interactions, these activities and interactions can be designed interestingly and meaningfully by applying learners' knowledge in authentic contexts, especially with the help of ubiquitous apps. Hence, designed learning activities can be effective in improving learning performance. Accordingly, the difference in performance between learners who used the ubiquitous app in authentic contexts and learners who used pencil/paper to do measurement tasks should be investigated. In addition, the correlation between performance and learning behaviors in peer assessment should also be explored to proffer the statistical evidence of the usefulness of peer assessment and peer comments in the educational practice.

Regarding learning behaviors, past studies mentioned that the indicators of learning behaviors include completing the tasks, sharing, and explaining ideas to others (Coolahan et al., 2000; Fredricks et al., 2016). In the present study, learning behaviors are considered necessary for learning, which can be measured by relying on the learning effectiveness and quantity of learning behaviors. The learning effectiveness of learners in authentic activities should be grounded on the scoring criteria of the completed tasks (Lindsay & Pamela, 2001; Tan & Hew, 2016). Learning behaviors in authentic activities is based on a real-life situation, such as measuring angles and length of geometry objects in surroundings. Meanwhile, activities recorded in the learning management systems (LMSs) are implied as to the quantity of learning behaviors that relates to the number of their frequency (Tan & Hew, 2016). Nevertheless, there are few studies available in the literature that combine both learning effectiveness and quantity of learning as an essential part of learning activities while exploring authentic contexts using ubiquitous apps.

The learning effectiveness and quantity of learning behaviors of learners are influenced by how the learning in authentic contexts is designed. In this regard, a hierarchical model of Bloom's taxonomy can be utilized to design learning activities and tasks (Anderson et al., 2001) based on authentic contexts. The first three levels in this taxonomy (i.e., remembering, understanding, and applying) are elaborated by doing some activities relating learners' mathematics knowledge to authentic contexts, such as measuring geometry objects in surroundings and making annotations. In peer assessment, analyzing and evaluating levels are carried out while the learner compares his/her work with others'. Learners can build new meaningful ideas by doing more experiences in authentic contexts to stimulate imagination and do a variety of creative measurements. To determine whether the learning effectiveness or quantity of learning in authentic contexts affects achievement, researchers need to evaluate the influence of learning effectiveness and quantity of learning on achievement by doing correlation and regression analysis. In addition, to know the effect of the designed learning on learners' cognitive abilities, it is needed to investigate not only the influence of the learners' learning effectiveness and quantity of learning to their cognitive abilities but also their perceptions toward the learning design.

Therefore, we recorded learners' learning behaviors in authentic measurement to get complete information on learning effectiveness and quantity of behaviors when learners engage in the ubiquitous learning environment (ULE). Ubiquitous Geometry (UG), a mobile android application, was designed and developed to facilitate learning angle and polygon concepts with authentic measurement and record their learning behaviors, including measurement, annotation, and peer assessment activities. As such, five research questions are addressed below.

- Is there any different learning performance between learners who use UG to do angle measurement tasks in authentic contexts and those who use protractors and pencil/paper to do such tasks?
- When learners engage in authentic learning using UG, what are the correlations between learning effectiveness, quantity of learning, and learning achievement?
- What is the prominent learning effectiveness and quantity of learning that can predict learners' learning achievement who engage in authentic learning using UG?
- What are the learning effectiveness and quantity of learning of learners that influence different cognitive abilities?
- What are the learners' perceptions of UG and their motivation for geometry learning in authentic contexts?



## **2. Literature review**

### **2.1. Learning activities in authentic contexts**

In enactivism, learning is a complex activity, which requires harmony between cognitive, physical, and environmental aspects. Instead, of mastering knowledge or abilities, a complex process that includes understanding, abstracting, and applying becomes the way how cognition and learning environment enact with each other. Enactivism paradigm emphasizes that embodiment and action can influence learners' cognition (Li, Clark, & Winchester, 2010) and become more popular in the interaction design and technology field to help learners create their individual learning environment (Winn, 2006). According to this idea, learning in authentic contexts is an educational implication that merges three essential aspects of enactivism, i.e., cognition, physical activities, and rich contexts (environmental aspect), into the learning activities. Past studies (Crompton, Burke, & Lin, 2019; Ekren & Keskin, 2017) have used the revised Bloom's taxonomy as a framework to examine the processes that took place in learning, which was supported by educational technologies. The technologies can help learners to overcome their difficulties (Hwang, Tsai, & Yang, 2008) while they do math tasks, e.g., angle and polygon measurement, in authentic contexts.

In addition, social interaction based on social constructivism theory has an essential role in cognitive development (Vygotsky, 1978). Vygotsky (1978) claimed that there is a distance between the knowledge developed individually and that developed through interacting with others. This notion was widely used in past studies to design the activities supported by technologies that help learners to do more interaction with others (Amory, 2018; Barbosa, Barbosa, & Rabello, 2016). The implementation of this notion in learning is that a learning process can be supported by a technology that could connect each individual with others via mutual observation, sharing, negotiation, and evaluation of problems (Clements & Battista, 1990). Following Hwang and Hu (2013), we designed peer assessment and comment activities as part of social constructivism that facilitated the interaction and communication with peers when learners learned in authentic contexts.

Enactivism and social constructivism were merged with revised Bloom's taxonomy (Anderson et al., 2001) into the design of learning activities and tasks. We utilized the first five cognitive levels of Bloom's taxonomy. The first three levels in the taxonomy, i.e., remembering, understanding, and applying, became our major focus. Learners should master the three levels while they did tasks individually. The use of authentic contexts would help learners to understand the concepts of mathematics more meaningfully by making daily life applications. After learners worked individually, they could compare their works with those of others in peer assessment, which means they would intend to do activities representing higher levels of cognition in Bloom's taxonomy, e.g., analyzing and evaluating. Learners would analyze and evaluate their works by comparing those with their peers'. They can make a new idea to solve tasks in wider authentic contexts. Moreover, the various experiences in authentic contexts can stimulate learners to draw a shape with specific criteria in its angle. It is useful to promote high cognitive levels in the taxonomy. Therefore, we also used Bloom's taxonomy to design pre-test/post-test and to evaluate cognitive levels based on learners' achievement.

### **2.2. Enhancing geometry learning with ubiquitous technology in authentic contexts**

Learners could discover their knowledge through interaction with environments and could apply it in different conditions (Purba et al., 2019). Learners will receive the knowledge and apply it in a real-life situation which useful to enhance learners' cognitive level. Thus, it was not surprising that designing activities based on learners' daily life would give more benefit to their learning outcomes (Hwang et al., 2015; Hwang et al., 2019).

In the past decade, several studies reported that the use of technologies in learning would support better learning outcomes. Geometer's Sketchpad (Erbaş & Yenmez, 2011) and CABRI (Bokosmaty, Mavilidi, & Paas, 2017) are computer-based platforms that are effective in exploring geometry objects to support higher-order thinking and manipulation ability of learners. However, these technologies can only be used in the classroom. They cannot support learning in authentic contexts and have limitations for tracking the learning process.

Recently, many studies have used mobile applications to support learning in authentic contexts. Particularly, the multimedia and portability of mobile devices can provide multiple representations with the tracking learning experience and allow the learner to learn anytime and anywhere (Hwang, Tsai, & Yang, 2008). These can be used to support learning in authentic contexts (Coffland & Xie, 2015). In addition, mobile devices can offer interactive representations that encourage learners' cognitive processes on concrete, visual, and abstract stages (Volk et al., 2017); thereafter, mobile devices will also improve learners' understanding of geometry concepts.

In the previous studies, the experiments were conducted to investigate the effectiveness of UG in learning the perimeter and area of two-dimensional shapes (Hwang et al., 2019; Hwang, Hoang, & Tu, 2020). The results revealed that UG was beneficial to enhance estimation ability, achievement, spatial ability, and geometry problem solving (Hwang et al., 2019; Hwang, Hoang, & Tu, 2020). However, these studies did not deeply address the relationship of learning achievement with learning behaviors (e.g., authentic measurements and annotations) and social interaction (e.g., peer assessment) in authentic contexts. Hence, we designed learning activities (e.g., authentic measurements, annotations, and peer assessment) to enhance the understanding of geometric concepts (Vitale, Swart, & Black, 2014) through authentic manipulation and measurement (Bokosmaty, Mavilidi, & Paas, 2017). Learners will use UG in tablet devices to do measurements in authentic contexts (authentic measurements); thereafter, they will be led to make annotations and do peer assessment. Accordingly, modified UG was required to facilitate learners doing object (in real-world) manipulation with multiple representations of geometric objects in tablet devices (artifact in virtual space). They also needed support to interact with their peers (social human). These three interactions in UG would be involved in four dimension spaces of the ubiquitous learning environment (ULE): real world, virtual space, personal space, and shared space (Li et al., 2004).

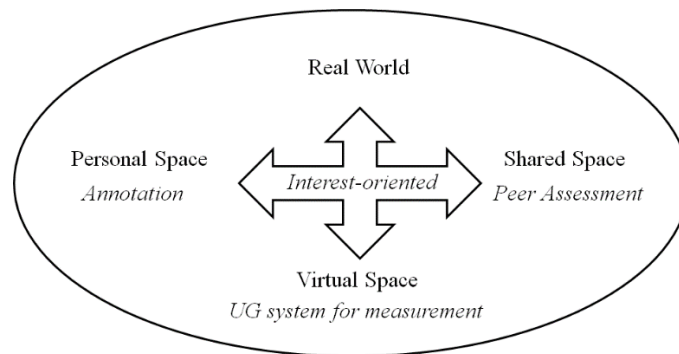


Figure 1. Framework of interest-oriented in the ubiquitous learning environment and learning experience with UG

### 2.3. Learning behaviors with measuring geometry in authentic contexts

Geometry learning is strongly related to geometry measurement. In geometry, the measurement of real objects, known as authentic measurement, can motivate and increase learning experience and enhance achievement (Hwang et al., 2019), especially for elementary school learners. Regarding geometry measurement in authentic contexts, learning behaviors need to be investigated deeply to know its influence on geometry learning.

With the help of the LMS (Wang, 2017), it is easier to measure and collect the number of activities in learning supported by mobile devices. The study (Rafaeli & Ravid, 1997) regarding the evaluation of learners' behaviors using learner logs indicated that there was a positive correlation between learners' achievement and their quantity of reading tasks. In this study, the quantity of learning behaviors was collected based on the framework in Figure 1, including three learning activities, namely, authentic measurements, annotations, and peer assessment. Authentic measurements comprised measuring angle and length of real geometric objects (MA), measuring elevation and depression angle (MED), and measuring polygon angle among different geographical places (MP), which are based on the angle, length, and polygon concepts. Regarding annotation activities, there were three types of annotations, including note drawing, note text, and note voice. In peer assessment, learners were able to give comments and responses to their peers by typing texts or drawing notes. Hence, we recorded comment drawing (CD), comment texts (CT), respond drawing (RD), and respond texts (RT).

### 3. Ubiquitous Geometry (UG) app supported by experience API

Experience API (xAPI), also known as Tin Can API and developed by Advanced Distributed Learning Initiative, is an open data interoperability specification originally developed to get a better picture of how, when, and why learning and performance happen both online and offline. xAPI system records data in a standardized format of xAPI statement, which is human and machine-readable. Afterward, the data are stored in the Learning Record Store (LRS) in an immutable format. This learning record has a powerful function in the case of tracing and recording learners' learning behaviors.

UG was designed and developed to help learners learn geometry concepts and record their learning behaviors. All data gathered in UG were stored online. Learning behavior data (xAPI statement) was stored in Yet LRS. Meanwhile, measurement, annotation, and peer assessment data were stored in google firebase. Therefore, learners could continue their work anytime and anywhere, as long as their devices were connected to the internet.

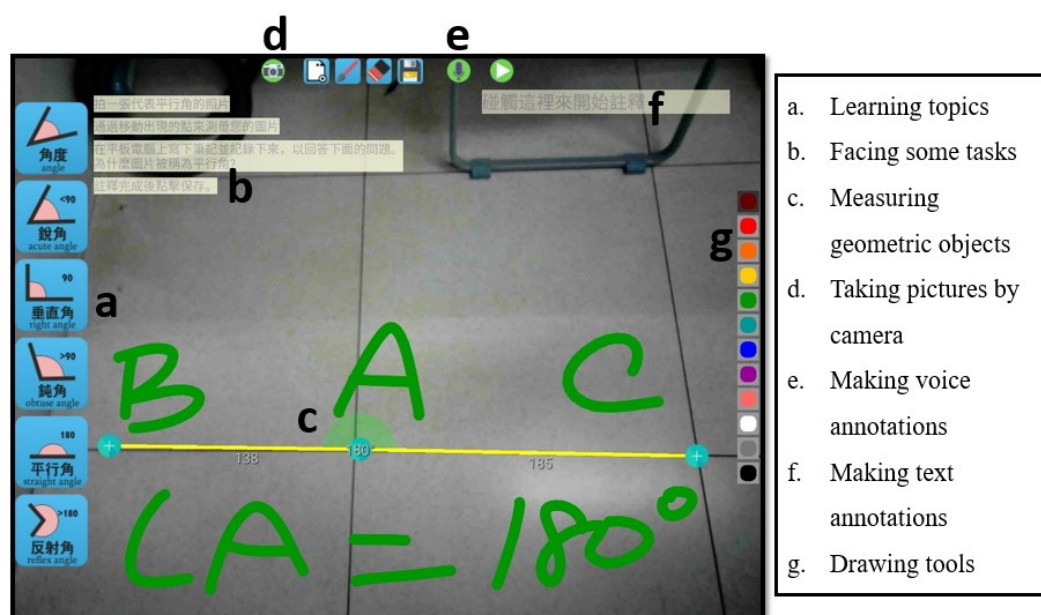


Figure 2. Preview of angle and polygon learning interface

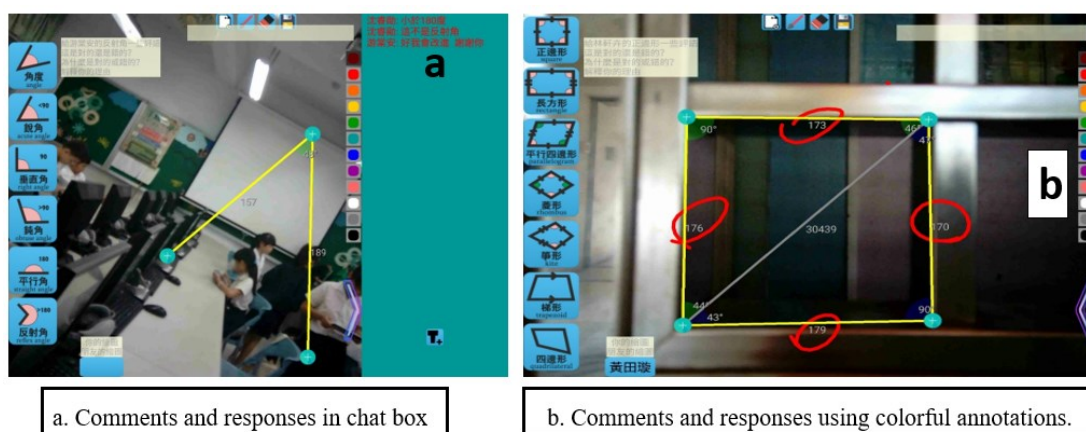


Figure 3. Preview of peer assessment interface

UG required learners to find real objects that represented geometric objects. Then, learners were asked to measure the angle and length of real-objects and to make annotations (see Figure 2). Afterward, learners were also asked to give their comments on peer assessment activity (see Figure 3).

## 4. Methods

### 4.1. Participants and experimental procedure

Participants were 53 fourth-grade learners (who were 9–10 years old) from an elementary school in North Taiwan. We had a successful collaboration with this school using information and communication technology (ICT) to enhance the learning and ICT literacy of its learners over ten years. Therefore, the participants were trained and became familiar with UG and the proposed learning activities, thereby having a good knowledge to do the learning activities. They were divided into two groups, namely, the experimental group (EG) and the control group (CG). The EG (26 learners) used UG, whereas the CG (27 learners) used protractors and pencil/paper to finish learning tasks. However, the two groups had the same learning materials and the same

instructor with more than 5 years of teaching experience in an elementary school. She had good experiences in teaching with technology; before the experiment, she had used UG and was familiar with it.

The experimental procedure is shown in Figure 4. Before we conducted the experiment, both groups were given a pre-test that aimed to know learners' prior knowledge. We conducted the experiment for a period of four weeks. First, learners in the EG were trained to be familiar with UG. Afterward, the EG were given tasks and allowed to use UG to explore their surroundings during break time and lunchtime. The use of UG was to help learners learn about angle and polygon concepts in mathematics. Conversely, learners in the CG were given protractors to do pencil/paper tasks as homework. According to the use of two different measurement tools, the design of the UG app required learners to do angle and polygon measurements of authentic objects in their surroundings and allowed them to make annotations, including voices, drawings, and texts, in their tasks and to write a comment to others' work immediately in real time. By contrast, the CG did the measurements using protractors to measure angles and polygons and wrote their results with texts and graphs in a paper-based way without authentic exploration and peer comment. This is because paper-based peer comment is not easy to conduct. In the end, we prepared a post-test for both groups and questionnaires and interviews for the EG.

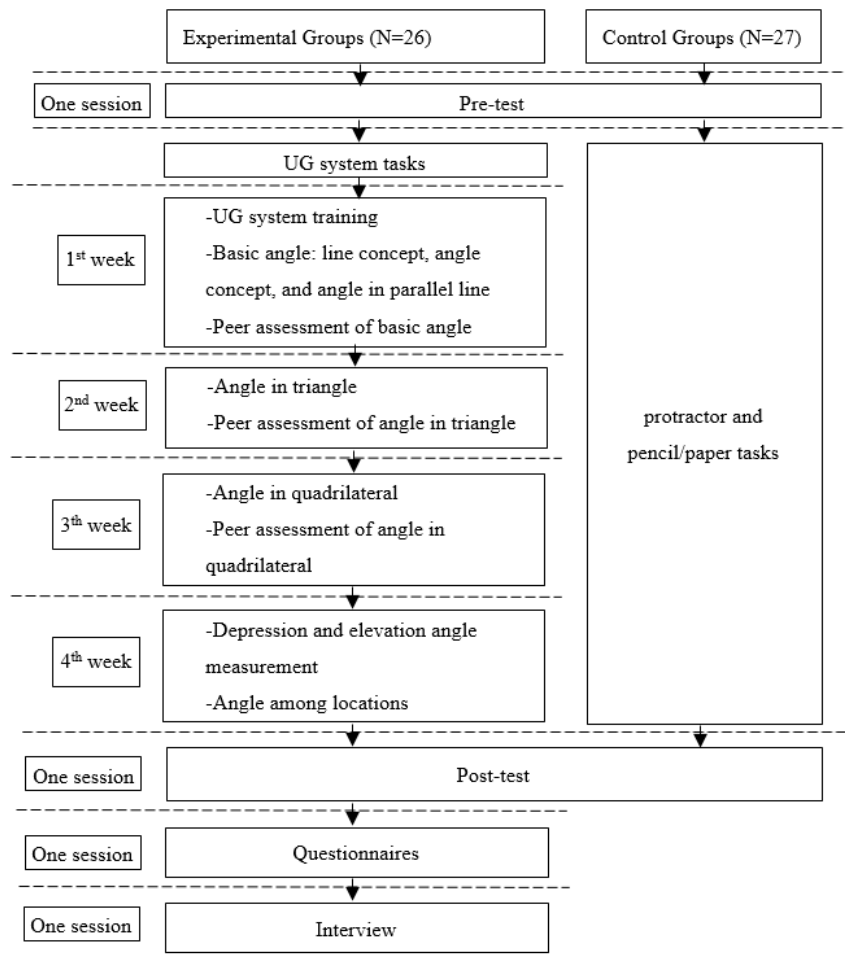


Figure 4. Experimental procedure

#### 4.3. Learning behaviors while using Ubiquitous Geometry

In the EG, we used two sets of data to know the learning behaviors in authentic measurements, annotations, and peer assessment. The first data were taken from the quantity of data recorded by xAPI based on their behaviors when using UG, which means how many times learners would do a particular activity. Besides the quantity of learning activities, the effectiveness of learning behaviors is also considered. This learning effectiveness can be used to evaluate the correctness of measurements, annotations, and peer assessment that had been done by learners. Thus, we consider learning behaviors with geometry measurement in quantity and learning effectiveness aspects. The quantity of learning is related to how many times learners measure authentic objects, make annotations, and do peer assessment in authentic contexts. Conversely, the learning effectiveness considers

the correctness of the three mentioned activities, which were scored by a mathematics teacher based on scoring criteria (see Appendix 1). Table 1 provides a detailed explanation of learners' learning behaviors.

*Table 1. Learners learning behaviors*

	Learning activities	Learners activities	Explanation
Learning effectiveness	Authentic measurements	MA	The score of angle and length measurements
		MED	The score of elevation and depression angle measurements
		MP	The score of angle in polygon measurements among different geographical places
	Annotations	ND	The score of note drawing
		NV	The score of note voice
		NT	The score of note text
	Peer assessment	CD	The score of comment drawing
		CT	The score of comment text
		RD	The score of respond drawing
		RT	The score of respond text
Quantity of learning	Authentic measurements quantity	qMA	Quantity of angle and length measurements
		qMED	Quantity of elevation and depression angle measurements
		qMP	Quantity of angle in polygon measurements among different geographical places
	Annotations quantity	qND	Quantity of note drawing
		qNV	Quantity of note voice
		qNT	Quantity of note text
	Peer assessment quantity	qCD	Quantity of comment drawing
		qCT	Quantity of comment text
		qRD	Quantity of respond drawing
		qRT	Quantity of respond text

#### 4.4. Research tools

##### 4.4.1. Questionnaires

Two questionnaires were used to discover how learners felt when using the system and after using the system in an authentic learning environment. The technology acceptance model (TAM) questionnaire based on past studies (Hwang, Hoang, & Tu, 2020; Purba et al., 2019) was used, and four dimensions (i.e., perceived usefulness (three items), perceived ease of use (two items), attitude toward use (three items), and behavioral intention (two items)) were investigated. In addition, the attention, relevance, confidence, satisfaction (ARCS) questionnaire based on past studies (Hwang, Hoang, & Tu, 2020; Purba et al., 2019) was utilized to know about learners' motivation in their learning activities using UG in the proposed learning environment. ARCS questionnaire consists of 14 items that represent four dimensions: attention (four items), relevance (four items), confidence (three items), and satisfaction (three items). This study used a five-point Likert scale with a starting point for "strongly agree" (5) and enclosed by "strongly disagree" (1).

##### 4.4.2. Pre-test and post-test

The pre-test and post-test comprise twelve angle and polygon concept questions (Q1–Q12) that were designed based on the first five levels of cognitive domain taxonomy (i.e., remembering, understanding, applying, analyzing, and evaluating) (Hwang et al., 2007; Kastberg, 2003; Žilková, Guncaga, & Kopáková, 2015). We discussed and designed the tests together with two experienced mathematics teachers who helped to evaluate the validity and reliability of tests. The first nine questions (Q1–Q9) are multiple-choice questions. These have 0 points (false) and 1 point (correct). Then, three questions (Q10–Q12) are essay questions. This part uses three kinds of evaluation in scoring, 0, 1, and 2 points. In the second part, learners would get the maximum point (2 points) if their answer was totally correct. If the answer were partially correct (the answer not complete or having a misconception), they would get 1 point. Learners would get 0 points if they could not answer, or their answer was totally wrong. Therefore, the total scores of these 12 questions were 15, which were normalized to 100.



Considering learners' cognitive abilities (Forehand, 2010), we used the same instrument (post-test), but we divided it into two group questions. Q1 to Q9 belong to low cognitive ability questions (remembering, understanding, and applying abilities), whereas Q10 to Q12 belong to high cognitive ability questions (analyzing and evaluating). The example of pre-test and post-test items can be found in Appendix 2.

#### 4.4.3. Interviews

Interviews were applied to explore what learners perceived when using UG for learning angle measurements in authentic contexts. Accordingly, three learners were selected to be interviewees based on their scores in the post-test (a student with high achievement, a student with middle achievement, and a student with low achievement). We also prepared open-ended questions, and all audio-recorded contents were analyzed to give an in-depth understanding of statistical results. These questions are as follows:

- Do you like using this system? Could you tell me which part of the system that you like or dislike?
- In your opinion, what learning activities do you think are important? Why?

#### 4.5. Data analysis

Four statistical analyses were conducted using IBM SPSS 20. First, an analysis of variance (ANOVA) was used to identify the equivalent of learners' prior knowledge before participating in the experiment. Second, an analysis of covariance (ANCOVA) was used to identify the differences in learning achievement between the EG and the CG. Third, Pearson correlation analysis was utilized to examine the correlations of learning behaviors with learning achievement. The last was the multiple regression analysis, used to identify prominent learning behaviors predicting learners' learning achievement who engage in authentic learning using UG. Moreover, the collected data from questionnaires were analyzed using Cronbach's alpha reliability test and descriptive analysis.

### 5. Results and discussions

#### 5.1. Learning achievement

The pre-test was used to know learners' prior knowledge related to angle and polygon measurements. Before doing ANCOVA, prior knowledge in the two groups and homogeneity of variance should be examined. Basically, based on ANOVA, there was no significant difference in learners' pre-test scores ( $F(1, 51) = 0.002, p = .969$ ) between these two groups (EG ( $M = 53.30, SD = 18.54$ ); CG ( $M = 53.09, SD = 20.79$ )). Levene's test indicates that the variance of pretest ( $F(1, 51) = 0.003, p = .958$ ) and learning achievement ( $F(1, 51) = .817, p = .370$ ) for learners in the EG and the CG are equal.

Table 2. The ANCOVA results for learning achievement considering pre-test scores as the covariate

Group (N)	Mean	SD	Adj. Mean	F	Sig.	$\eta^2$
Experimental group (N = 26)	75.128	17.844	75.063	5.190*	0.027	0.094
Control group (N = 27)	62.221	23.019	62.206			

Note. R Squared = 0.117; \* $p < .05$ .

In Table 2, the results of the ANCOVA test show that there is a statistically significant difference in achievement between these two groups ( $F(1, 50) = 5.190, p = .027$ ). This result indicates that UG is beneficial for learners' geometry learning; hence, learners in the EG can better understand and solve problems concerning angles and polygons than those in the CG.

UG was designed to be used in tablet devices to support learners in measuring angles and length of real objects, making annotations, and writing comments in peer assessment. Following the enactivism paradigm, these three learning activities give highly positive benefits for learners' learning performance because their activities are inseparable from authentic contexts. Learners were encouraged to apply geometry concepts in real practice by measuring the angles of various real objects in their surroundings. This activity has a positive correlation with learners' geometry learning achievement and their geometry thinking ability (Hwang et al., 2015; Hwang et al., 2019). Through the UG support, learners receive the concept in abstract information and represent their image concept in a real situation.

In addition, tablet devices with multimedia and multiple sensory interactions could support learning with multiple representations, including voices, drawings, texts, and real objects, that facilitate learning outcomes in cognitive, affective, and psychomotor learning domains (Volk et al., 2017). As such, UG is equipped with a feature that enables learners to make multiple representations in their annotations. Annotations can represent learners' understanding of geometry concepts, so those annotations have important roles in increasing learning achievement (Hwang et al., 2011). In terms of social constructivists, learners could review their peer annotation, which could increase their understanding of mathematics concepts (Hwang et al., 2011). Moreover, social interaction while doing peer assessment can enhance their understanding and give them a chance to communicate their idea to others during the process.

Two confounding variables were related to activity design in this study, namely, authentic exploration and peer comment. However, this study did not elucidate whether the difference in learning achievement between the EG and the CG was influenced by authentic exploration or peer comment. Therefore, we will address this issue in our future experiment.

Table 3. Pearson correlation between learning effectiveness and learning achievement

Var.	LA	MA	MED	MP	ND	NV	NT	CD	CT	RD	RT
LA	1										
MA	0.393*	1									
MED	0.579**	0.177	1								
MP	0.059	0.517**	0.000	1							
ND	0.467*	0.656**	0.368	0.310	1						
NV	0.304	-0.062	0.390*	0.216	0.327	1					
NT	0.127	-0.189	0.114	-0.284	0.090	0.215	1				
CD	0.467*	0.256	0.061	0.073	0.420*	0.103	-0.189	1			
CT	0.311	0.711**	0.154	0.352	0.645**	-0.038	-0.093	0.364	1		
RD	0.353	0.436	0.320	0.217	0.632**	0.068	0.099	0.451*	0.587**	1	
RT	0.241	0.570**	0.075	0.176	0.434*	-0.152	0.021	0.180	0.465*	0.129	1

Note. \* $p < .05$ ; \*\* $p < .001$ , LA = learning achievement.

## 5.2. Correlation among learning effectiveness, quantity of learning, and learning achievement

According to learning behaviors, UG recorded all of learners' activities during the experimental period. Hence, their learning effectiveness can be identified by scoring the outcomes in each activity. As shown in Table 3, learning achievement positively correlated with MA ( $r = 0.393$ ,  $p = .047$ ), MED ( $r = 0.579$ ,  $p = .002$ ), ND ( $r = 0.467$ ,  $p = .016$ ), and NV ( $r = 0.304$ ,  $p = .049$ ). MA positively correlated with MP ( $r = 0.517$ ,  $p = .007$ ), ND ( $r = 0.656$ ,  $p = .000$ ), CT ( $r = 0.711$ ,  $p = .000$ ), RD ( $r = 0.436$ ,  $p = .026$ ), and RT ( $r = 0.570$ ,  $p = .002$ ). MED positively correlated with NV ( $r = 0.390$ ,  $p = .049$ ). ND positively correlated with CD ( $r = 0.420$ ,  $p = .003$ ), CT ( $r = 0.645$ ,  $p = .000$ ), RD ( $r = 0.632$ ,  $p = .001$ ), and RT ( $r = 0.434$ ,  $p = .027$ ). CD positively correlated with RD ( $r = 0.451$ ,  $p = .021$ ). CT positively correlated with RD ( $r = 0.587$ ,  $p = .002$ ), and RT ( $r = 0.465$ ,  $p = .017$ ).

Table 4. Pearson correlation between quantity of learning and learning achievement

Var.	Post-test	qMA	qMED	qMP	qND	qNV	qNT	qCD	qCT	qRD	qRT
Post-test	1										
qMA	0.068	1									
qMED	0.579**	0.156	1								
qMP	0.152	0.143	0.118	1							
qND	0.397*	0.172	0.221	0.303	1						
qNV	-0.105	-0.263	0.090	0.160	0.437*	1					
qNT	0.073	-0.036	0.103	-0.003	0.487*	0.305	1				
qCD	0.404*	-0.038	0.047	0.101	0.526**	0.245	0.072	1			
qCT	0.309	-0.03**	0.005	0.393*	0.381	0.312	0.365	0.237	1		
qRD	0.272	-0.030	-0.101	-0.234	0.081	-0.367	-0.066	0.192	0.187	1	
qRT	0.191	0.113	0.026	0.352	0.508**	0.224	0.720**	0.162	0.498**	-0.108	1

Note. \* $p < .05$ ; \*\* $p < .001$ .

On the other hand, the quantity of learning based on log file data was also analyzed. As shown in Table 4, learning achievement positively correlated with qMED ( $r = 0.579, p = .002$ ), qND ( $r = 0.397, p = .045$ ), and qND ( $r = 0.404, p = .041$ ). qMP positively correlated with qCT ( $r = 0.393, p = .047$ ). Besides learning achievement, qND positively correlated with qNV ( $r = 0.437, p = .026$ ), qNT ( $r = 0.487, p = .012$ ), qCD ( $r = 0.526, p = .006$ ), and qRT ( $r = 0.508, p = .008$ ). qNT positively correlated with qRT ( $r = 0.720, p = .000$ ). qCT positively correlated with qRT ( $r = 0.498, p = .010$ ).

A significant correlation is shown between achievement (post-test) and MED in both learning effectiveness and quantity of learning behaviors. This implied that measuring real objects was a good activity that could support achievement. In MED, learners were allowed to move from one place to another to explore their understanding of the basic concept of angle measurements. In this activity, learners were excited, and they only focused on the simple angle measurements.

In annotation activities, ND is beneficial for learning achievement in both learning effectiveness and the quantity of learning. In ND activity, learners could use multiple representations such as figures, texts, and symbols in their notes. Furthermore, they could use different colors depending on their interest. This activity could give them a chance to practice their ability to make annotations interestingly.

From peer assessment, CD and qCD were beneficial for achievement. The result also showed that peer assessment behaviors were correlated each other in learning engagement RD to CD ( $r = 0.451, p = .021$ ), RT to CT ( $r = 0.465, p = .017$ ), and RD to CT ( $r = 0.587, p = .002$ ). Unfortunately, most of the learners' CT were "good job" and "you are wrong," which were not followed by reasons for their comments. These kinds of comments are not beneficial for improving understanding, so that becomes a possible reason why only CD correlated with the post-test.

### 5.3. Multiple regression of variables in quantity of learning and learning effectiveness toward learning achievement

Based on multiple regression analysis results, the predictor variable of the quantity of learning that has the most influence on learning achievement is qMED ( $M = 7.00, SD = 3.175, B = 3.155, p = .001$ ; see Table 5). Similarly, in learning effectiveness, the predictor variable that would give the most influence on learning achievement is MED ( $M = 7.00, SD = 3.175, B = 3.106, p = .001$ ; see Table 6). Based on descriptive statistics, all measurements in elevation and depression angles were correct. Such findings could be caused by the fact that this activity was new and used learners' knowledge in simple angle measurements. Learners were able to measure an angle between two objects on the basis of their eyes' viewpoint as the central point. They not only use real objects to apply their knowledge but also use parts of their body, such as their eyes and hands, to measure elevation and depression angles. Therefore, learners felt excited and could further explore their understanding of the basic concepts of angle measurements in a simple object. This finding strengthens those of previous studies (Morris, Finnegan, & Wu, 2005; Wang, 2017), which claimed that learning interaction and learning engagement could affect learning achievement. Regarding the embodiment concept in enactivism, the MED activity encouraged learners to interact in authentic contexts and involve their sensory and motor processes to support cognitive development (Li, Clark, & Winchester, 2010). Thus, teachers need to design such learning activities that can embrace both sensory and motor interaction into learning in authentic contexts.

Table 5. Regression model summary of variables in quantity of learning toward learning achievement

Model	Unstandardized coefficients		Standardized coefficients	<i>t</i>	Sig.
	B	Std. error	Beta		
(Constant)	48.009	6.718		7.146	0.000
qMED	3.155	0.848	0.561	3.720	0.001
qCD	0.609	0.243	0.378	2.503	.020

Note.  $R^2 = 0.477$ , adjusted  $R^2 = 0.432$ .

In the second place, learning achievement can also be predicted by both qCD ( $M = 8.269, SD = 11.069, B = 0.609, p = .02$ ) and CD ( $M = 6.846, SD = 6.017, B = 1.285, p = .006$ ). Based on descriptive statistics, the number of CD cannot be used as a good representation of data ( $SD > M$ ). This result shows that the effectiveness of CD has an important role in enhancing learning achievement. Learners could share their knowledge while drawing comments. At the same time, they also could reflect on others' work and compare it with theirs. Learners not only criticized peers' solutions but also helped them to rearrange their solutions to make a good answer and

improve their understanding (Hwang & Hu, 2013). Moreover, the use of different colors in the CD makes it easier and clearer for learners to write their comments because they drew directly in peers' solutions without typing texts that need more time.

*Table 6. Regression model summary of variables in learning effectiveness toward learning achievement*

Model	Unstandardized coefficients		Standardized coefficients	t	Sig.
	B	Std. error	Beta		
(Constant)	44.585	6.722		6.632	0.000
MED	3.106	0.811	0.553	3.828	0.001
CD	1.285	0.428	0.434	3.003	0.006

Note.  $R^2 = 0.522$ , adjusted  $R^2 = 0.481$ .

*Table 7. Pearson correlation between learning effectiveness and learning achievement*

Learning effectiveness	Low cognitive ability	High cognitive ability
MA	0.353	0.421*
MED	0.553**	0.439*
MP	-0.012	0.238
ND	0.234	0.601**
NV	0.168	0.356
NT	0.195	0.107
CD	0.378	0.347
CT	0.277	0.342
RD	0.252	0.350
RT	0.156	0.309

Note. \* $p < .05$ ; \*\* $p < .001$ .

#### 5.4. Correlation between learning effectiveness and cognitive abilities

Considering learners' cognitive abilities, we also investigated the relationship between learners' learning effectiveness and cognitive abilities using Pearson correlation. As shown in Table 7, low cognitive ability had significant positive correlations only with MED ( $r = 0.553$ ,  $p = .003$ ). On the other hand, high cognitive ability had significant positive correlations with three variables: MA ( $r = 0.421$ ,  $p = .032$ ), MED ( $r = 0.439$ ,  $p = .025$ ), and ND ( $r = 0.601$ ,  $p = .001$ ). It was meant that low cognitive ability could obtain positive benefits by doing MED. From the enactivism perspective, the use of particular tasks, e.g., measuring real objects, will influence learners' effective behaviors (Lozano, 2017). These effective behaviors will sustain individual motivation to continue learning in authentic contexts (Simmt & Kieren, 2015). The following opinions were found in the interview of learners with low achievement.

S1: "I think measuring elevation and depression angles is important because it is practical."

The high cognitive ability could obtain positive benefits by measuring the angle in authentic contexts, especially in MA and MED. When learners measured angles and lengths of objects in surroundings using UG, they could explore their knowledge by doing the first three cognitive learning activities, including remembering, understanding, and applying particular geometry concepts in authentic contexts. Learners could also make a good note drawing based on their understanding of angle concepts by analyzing and evaluating the picture of the measured object that implied high cognitive activities in Bloom's taxonomy. Furthermore, in this study, the imagination belonging to the high cognitive ability was possibly stimulated by measuring and drawing a shape with specific criteria in its angles, e.g., a triangle with one angle is an obtuse angle. After completing authentic measurement and peer activities, learners could shape their own learning experiences by comparing with peers' work and get new ideas to make their own conclusion. Learners' conclusions can also make them internalize the concepts used to measure single angles with different criteria in authentic contexts, such as MA and MED, thereby increasing learning achievement in drawing a shape with specific angles. The following opinions were found in the interview of learners with high achievement.

S2: "I like using the system and measuring the angle of the real object because it makes me more understand about kinds of angle."

S3: "I think drawing a note on my work is important because I can understand my work."

Based on the observation of learning activities, in MA, learners applied concepts of single angle in different characteristics by measuring real objects, such as door corners and other objects. In MED, learners imagined the single angle formed while they moved their eyes from a horizontal position to look at the objects up or bottom. By doing MED activity, learners could widen their knowledge of angle concepts application in the abstract space by imagining lines and corners. Thus, MED had contributed to connect the abstract geometry concepts with real applications in physical worlds and enrich learners' embodied experiences. However, a further in-deep investigation was required in future studies to reconfirm the relationship between such activities (MA and MED) and cognitive abilities and the reasons behind them.

The aforementioned findings imply the empirical evidence that the cognitive abilities based on Bloom's taxonomy framework could be used to identify kinds of learning behaviors enhancing geometry learning in authentic contexts.

### **5.5. Perception and motivation toward learning experience using Ubiquitous Geometry app**

Learners' perception and motivation data were collected using TAM and ARCS questionnaires, respectively. The reliability of both questionnaires was tested using Cronbach's alpha test. The results indicate that both questionnaires, TAM (Cronbach's alpha = 0.92) and ARCS (Cronbach's alpha = 0.70), have good reliability and are categorized as acceptable constructs (Hair et al., 2010). TAM questionnaire resulted that most learners scored high for all items. In detail, the means scores are 4.11 for perceived usefulness, 3.67 for perceived ease of use, 3.90 for attitude toward use, and 3.62 for behavioral intention toward using UG. These results indicate that learners have a positive perception of the use of UG while learning in authentic contexts. Furthermore, most of them intend to use UG in the future.

The ARCS questionnaire results show a partially high degree of learners' learning motivation. The mean score of attention, relevance, and satisfaction almost reached 4 points: 3.64, 3.88, and 3.69, respectively. In addition, the mean score of confidence was 3, which implies that learners did not have high confidence toward a used system in an authentic learning environment. According to the learners' perspective, they had difficulties in measuring the real object because, for the first time using UG, they could not find geometry objects in their surroundings. Moreover, learners with low achievement felt confused when they did peer activities (giving comments and responding to the other learners' work). Consequently, it could reduce their confidence while using UG for peer assessment.

## **6. Conclusions**

The study reveals several important findings. First, learners who use UG significantly outperformed those who use protractors and pencil/paper. Regarding the further analysis of learning behaviors in EG, earlier studies have emphasized the learning behaviors and achievement in learning with UG (Hwang et al., 2019; Hwang, Hoang, & Tu, 2020). However, the previous studies focused on the quantity of learning behaviors (Hwang et al., 2019) and problem-solving (Hwang, Hoang, & Tu, 2020) to predict estimation and geometry abilities. In this study, measuring objects in authentic contexts, making annotations, and assessing peers' works highly affect learning achievement and help cognitive development. A possible reason is that, based on enactivism theory, learners enact geometry knowledge and real-life application by doing authentic measurements (Hwang et al., 2019; Hwang, Hoang, & Tu, 2020) and making annotations. Moreover, giving comments with stimuli from multiple representations of authentic contexts in peer assessment is very helpful in enhancing the experiences and effectiveness of learning in authentic contexts (Hwang & Hu, 2013). This is because new knowledge can be developed by interacting with others (Chung, Hwang, & Lai, 2019; Engeström, 1999; Vygotsky, 1978). Second, in the case of the correlation with learning achievement, both learning effectiveness and quantity of learning indicate a similar result that MED activity was the most influential engagement to the learning achievement of the EG. The effectiveness of comment drawing had an important role in improving the learning achievement of EG. Third, learners in the EG with low cognitive abilities were only influenced by MED; those with high cognitive abilities were influenced by MA, MED, and ND. Measuring different angles and lengths of real objects can help learners to understand geometry properties and related knowledge. Additionally, the learners of EG could understand their works by drawing notes in measurement pictures. Another finding is that learners have a good perception of UG (in terms of usability and ease of use) and have high enough motivation (attention, relevance, and satisfaction) in authentic learning.

However, we have several limitations in this study. First, this study could not clarify whether the difference in learning achievement between the EG and the CG was affected by authentic exploration or peer comments. Second, our experiment focuses on how the effectiveness of learning could influence learning performance, but we do not have further analysis on how it could influence cognitive engagements (including interest and strategies of learning). Therefore, in the future, we would like to expand our experimental design to investigate the influence of two different learning activities, i.e., authentic exploration and peer comment on learning achievement. Moreover, we would like to focus on learners' cognitive engagements.

## References

- Amory, A. (2018). Use of the collaboration-authentic learning-technology/tool mediation framework to address the theory-praxis gap. In T.-W. Chang, R. Huang, & Kinshuk (Eds.), *Authentic Learning Through Advances in Technologies* (pp. 61-73). Singapore: Springer Singapore.
- Anderson, L. W. (Ed.), Krathwohl, D. R. (Ed.), Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Rath, J., & Wittrock, M. C. (2001). *A Taxonomy for learning, teaching, and assessing: A Revision of bloom's taxonomy of educational objectives* (Complete ed.). New York, NY: Longman.
- Barbosa, J., Barbosa, D., & Rabello, S. (2016). A Collaborative model for ubiquitous learning environments. *International Journal on E-Learning*, 15(1), 5-25.
- Barak, M & Asakle, S. (2018). AugmentedWorld: Facilitating the creation of location-based questions. *Computers and Education*, 121, 89-99.
- Bokosmaty, S., Mavilidi, M.-F., & Paas, F. (2017). Making versus observing manipulations of geometric properties of triangles to learn geometry using dynamic geometry software. *Computers & Education*, 113, 313-326.
- Chung, C.-J., Hwang, G.-J., & Lai, C.-L. (2019). A Review of experimental mobile learning research in 2010–2016 based on the activity theory framework. *Computers & Education*, 129, 1-13.
- Clements, D. H., & Battista, M. T. (1990). Constructivist learning and teaching. *Arithmetic Teacher*, 38(1), 34-35.
- Coffland, D. A., & Xie, Y. (2015). The 21st century mathematics curriculum: A Technology enhanced experience. In X. Ge, D. Ifenthaler, & J. M. Spector (Eds.), *Emerging Technologies for STEAM Education* (pp. 311-329). Switzerland: Springer.
- Coolahan, K., Fantuzzo, J., Mendez, J., & McDermott, P. (2000). Preschool peer interactions and readiness to learn: Relationships between classroom peer play and learning behaviors and conduct. *Journal of Educational Psychology*, 92(3), 458-465.
- Crompton, H., Burke, D., & Lin, Y. C. (2019). Mobile learning and student cognition: A Systematic review of PK-12 research using Bloom's Taxonomy. *British Journal of Educational Technology*, 50(2), 684-701.
- Engeström, Y. (1999). Activity theory and individual and social transformation. In Y. Engeström, R. Miettinen, & R.-L. Punamäki (Eds.), *Perspectives on Activity Theory* (pp. 19–38). New York, NY: Cambridge University Press.
- Ekren, G., & Keskin, N. (2017). Using the revised bloom taxonomy in designing learning with mobile apps. *GLOKALde*, 3(1), 13–28.
- Erbas, A. K., & Yenmez, A. A. (2011). The Effect of inquiry-based explorations in a dynamic geometry environment on sixth grade students' achievements in polygons. *Computers & Education*, 57(4), 2462-2475.
- Ernest, P. (2010). Reflections on theories of learning. In B. Sriraman, & L. English (Eds.), *Theories of Mathematics Education* (pp. 39-47). Verlag Berlin Heidelberg, Germany: Springer.
- Forehand, M. (2010). Bloom's taxonomy. In M. Orey (Ed.), *Emerging perspectives on learning, teaching, and technology* (pp. 41-47). Zurich, Switzerland: Global Text.
- Fredricks, J. A., Wang, M.-T., Linn, J. S., Hofkens, T. L., Sung, H., Parr, A., & Allerton, J. (2016). Using qualitative methods to develop a survey measure of math and science engagement. *Learning and Instruction*, 43, 5-15.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). New Jersey: Pearson.
- Herrington, J., & Kervin, L. (2007). Authentic learning supported by technology: Ten suggestions and cases of integration in classrooms. *Educational Media International*, 44(3), 219-236.
- Herrington, J., Oliver, R., & Reeves, T. C. (2003). Patterns of engagement in authentic online learning environments. *Australasian Journal of Educational Technology*, 19(1), 59-71.
- Hwang, G.-J., Tsai, C.-C., & Yang, S. J. H. (2008). Criteria, strategies and research issues of context-aware ubiquitous learning. *Educational Technology & Society*, 11(2), 81-91.

- Hwang, W.-Y., Chen, N.-S., Dung, J.-J., & Yang, Y.-L. (2007). Multiple representation skills and creativity effects on mathematical problem solving using a multimedia whiteboard system. *Educational Technology & Society*, 10(2), 191-212.
- Hwang, W.-Y., Hoang A., & Tu, Y.-H. (2020). Exploring authentic contexts with ubiquitous geometry to facilitate elementary school students' geometry learning. *The Asia-Pacific Education Researcher*, 29, 269–283.
- Hwang, W.-Y. & Hu, S.-S. (2013). Analysis of peer learning behaviors using multiple representations in virtual reality and their impacts on geometry problem solving. *Computer & Education*, 62, 308-319.
- Hwang, W.-Y., Lin, L.-K., Ochirbat, A., Shih, T. K., & Kumara, W. (2015). Ubiquitous Geometry: Measuring authentic surroundings to support geometry learning of the sixth-grade students. *Journal of Educational Computing Research*, 52(1), 26-49.
- Hwang, W.-Y., Purba, S. W. D., Liu, Y.-F., Zhang, Y.-Y., & Chen, N.-S. (2019). An Investigation of the effects of measuring authentic contexts on geometry learning achievement. *IEEE Transactions on Learning Technologies*, 12(3), 291-302.
- Hwang, W. Y., Chen, N. S., Shadieff, R., & Li, J. S. (2011). Effects of reviewing annotations and homework solutions on math learning achievement. *British Journal of Educational Technology*, 42(6), 1016-1028.
- Kastberg, S. E. (2003). Using bloom's taxonomy as a framework for classroom assessment. *The Mathematics Teacher*, 96(6), 402-405.
- Kong, Q.-P., Wong, N.-Y, & Lam, C.-C. (2003). Student engagement in mathematics: Development of instrument and validation of construct. *Mathematics Education Research Journal*, 15(1), 4–21.
- Lai C.-L. & Hwang G.-J. (2015). An Interactive peer-assessment criteria development approach to improving students' art design performance using handheld devices. *Computers & Education*, 85, 149-159.
- Li, Q., Clark, B., & Winchester, I. (2010). Instructional design and technology grounded in enactivism: A Paradigm shift?. *British Journal of Educational Technology*, 41(3), 403-419.
- Li, L., Zheng, Y., Ogata, H., & Yano, Y. (2004, September). *A Framework of ubiquitous learning environment*. Paper presented at the 4th International Conference on Computer and Information Technology, Wuhan, China.
- Lindsay, C. & Pamela, R. A. (2001). Exploring the technical quality of using assignments and student work as indicators of classroom practice. *Educational Assessment*, 7(1), 39-59.
- Lozano, M.-D. (2017). Investigating task design, classroom culture and mathematics learning: An enactivist approach. *ZDM Mathematics Education*, 49, 895–907.
- Morris, L. V., Finnegan, C., & Wu, S.-S. (2005). Tracking student behavior, persistence, and achievement in online courses. *The Internet and Higher Education*, 8(3), 221-231.
- Nicaise, M., Gibney, T., & Crane, M. (2000). Toward an understanding of authentic learning: Student perceptions of an authentic classroom. *Journal of Science Education and Technology*, 9(1), 79-94.
- Purba, S. W. D., Hwang, W.-Y., Pao, S.-C., & Ma, Z.-H. (2019). Investigation of inquiry behaviors and learning achievement in authentic contexts with the ubiquitous-physics app. *Educational Technology & Society*, 22(4), 59-76.
- Rafaeli, S., & Ravid, G. (1997, October). *Online, web-based learning environment for an information systems course: Access logs, linearity and performance*. Paper presented at the Information Systems Education Conference, Orlando, Florida, USA.
- Simmt, E., & Kieren, T. (2015). Three “moves” in enactivist research: A reflection. *ZDM Mathematics Education*, 47(2), 307–317.
- Tan, M. & Hew K. F. (2016). Incorporating meaningful gamification in a blended learning research methods class: Examining student learning, engagement, and affective outcomes. *Australasian Journal of Educational Technology*, 32(5), 19-34.
- Vygotsky, L.S. (1978). *Mind in society: The Development of higher psycho-logical processes*. Cambridge, MA: Harvard University Press.
- Vitale, J. M., Swart, M. I., & Black, J. B. (2014). Integrating intuitive and novel grounded concepts in a dynamic geometry learning environment. *Computers & Education*, 72, 231-248.
- Volk, M., Cotič, M., Zajc, M., & Starcic, A. I. (2017). Tablet-based cross-curricular maths vs. traditional maths classroom practice for higher-order learning outcomes. *Computers & Education*, 114, 1-23.
- Wang, F. H. (2017). An Exploration of online behaviour engagement and achievement in flipped classroom supported by learning management system. *Computers & Education*, 114, 79-91.
- Wang, M.-T., Fredricks, J. A., Ye, F., Hofkens, T. L., & Linn, J. S. (2016). The Math and science engagement scales: Scale development, validation, and psychometric properties. *Learning and Instruction*, 43, 16-26.

Winn, W. (2006). Functional contextualism in context: A Reply to fox. *Educational Technology Research and Development*, 54(1), 55–59.

Žilková, K., Guncaga, J., & Kopácová, J. (2015). (Mis)conceptions about geometric shapes in pre-service primary teachers. *Acta Didactica Napocensia*, 8(1), 27-35.

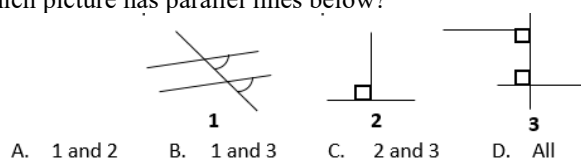
## Appendix 1. Scoring the learning effectiveness

Learning activities	Scores
MA, MED, and MP	1: measure the object correctly. 0: measure the object incorrectly.
ND, CD, and RD	4: draw mathematical principles correctly. 3: draw mathematical principles nearly correct. 2: draw mathematical principles incorrectly. 1: draw no meaningful figures. 0: draw nothing or irrelevant figures.
NT, CT, and RT	4: write mathematical principles correctly. 3: write mathematical principles nearly correct. 2: write mathematical principles incorrectly. 1: write no meaningful texts. 0: write nothing or irrelevant texts.
NV	4: record mathematical principles correctly. 3: record mathematical principles nearly correct. 2: record mathematical principles incorrectly. 1: record no meaningful voices. 0: record nothing or irrelevant voices.

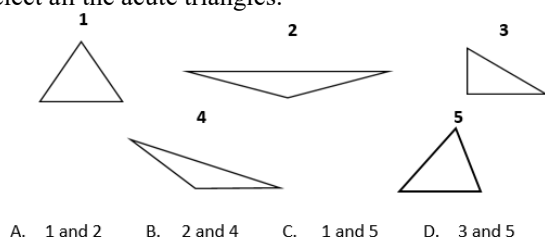
## Appendix 2. Examples of pre-test and examples of post-test

Examples of pre-test (Q1 and Q2):

Q1. Which picture has parallel lines below?



Q2. Select all the acute triangles.



Examples of post-test (Q12):

Q12. Draw a quadrilateral formed by two right triangles and an isosceles triangle. (Write the size of each angle of the triangle). Explain if you cannot draw it!



### Appendix 3. TAM and ARCS

TAM questionnaire items	Mean	SD
<b>Perceived Usefulness</b>	4.11	0.92
UG helps to improve my knowledge in learning Angle and Polygon.		
UG helps to improve my performance in learning Angle and Polygon.		
UG is effective for learning Angle and Polygon concepts.		
<b>Perceived Ease of Use</b>	3.67	1.00
It is easy for me to use UG.		
It is easy for me to understand UG.		
<b>Attitude Toward Use</b>	3.90	1.12
I believe that using UG is a good idea.		
I believe that using UG is advisable.		
I am satisfied in using UG.		
<b>Behavioral Intention</b>	3.62	1.01
I intend to use UG in the future.		
I will continue using UG increasingly in the future.		
<i>Note.</i> SD = standard deviation.		
ARCS questionnaire items	Mean	SD
<b>Attention</b>	3.64	1.03
It was interesting when I used UG to measure angles of objects in my surrounding.		
Taking pictures, making annotations and sound records helped to hold my attention.		
The learning activities by using UG could stimulate my curiosity.		
The repetition study with peer learning activities caused me to get bored sometimes.		
<b>Relevance</b>	3.88	0.92
It is clear to me how I used UG to learn concept of angle and polygon by using object in my surrounding.		
Measuring angle of object in my surrounding make me know how the concepts of angle were used in daily life.		
During the activities, I know the examples of used concept of angle and polygon in daily life.		
The way of learning concept of angle and polygon using UG was not relevant to my needs because I already knew most of it.		
<b>Confidence</b>	3.00	1.17
When I used UG at the first time to measure the angle, I had the impression that it would be easy for me.		
The learning activities by using UG were too difficult.		
After learning concept of angle and polygon using UG, I was confident that I would be able to pass a test on it.		
<b>Satisfaction</b>	3.69	1.12
Completing the activity measurement using UG gave me a satisfying feeling.		
I enjoyed using UG to explore concept of angle and polygon in my surrounding.		
The learning activities with peer helped me understand about concept angle and polygon.		

## The Impact of Game Playing on Students' Reasoning Ability, Varying According to Their Cognitive Style

Tsung-Yen Chuang<sup>1\*</sup>, Martin K.-C. Yeh<sup>2</sup> and Yu-Lun Lin<sup>1</sup>

<sup>1</sup>National University of Tainan, Taiwan // <sup>2</sup>Penn State University – Brandywine, Pennsylvania, USA //  
chuangyen@mail.nutn.edu.tw // martin.yeh@psu.edu // allenwubai@gmail.com

\*Corresponding author

(Submitted July 10, 2020; Revised September 4, 2020; Accepted February 25, 2021)

**ABSTRACT:** Students with different cognitive styles benefit from different instructional strategies, including learning through playing video games. Although playing video games can be an effective learning method, we do not know its impact on the reasoning ability of students with different cognitive styles. The purposes of this study are to investigate whether students with different cognitive styles improve their reasoning ability after playing video games and whether the effect is the same for all students. We used a pretest-posttest experimental design with multivariate analyses and found that elementary school students' reasoning ability improved reliably after playing a puzzle adventure game for four weeks, twice a week. In addition, field-independent students' reasoning ability improved reliably more than field-dependent students did. Students with different cognitive styles also demonstrated noticeably different information search strategies during game playing. Our work answers the questions regarding the impact of playing video games in students' reasoning ability and in students with different cognitive styles. We also suggested guidelines of designing educational video games for field-dependent and field-independent students. Future studies are needed to expand our understanding to the relationships between other types of video game, cognitive ability, and cognitive styles.

**Keywords:** Cognitive style, Digital game, Reasoning ability, Game-based learning

### 1. Introduction

Reasoning is a critical thinking skill that people frequently use in daily life. To solve problems, people need the reasoning ability to generate rules from a complex reality, to evaluate and judge relations from external information, and eventually, to produce a solution. While some studies showed that abstract thinking, the use of related knowledge, and inductive and deductive reasoning are important factors in enhancing student's learning process (Kline, 1994; Leighton & Sternberg, 2004; Nickerson, 1991), others had attempted to better understand individual differences in the human reasoning process (Carroll, 1993; Lohman & Lakin, 2011; Sternberg, 1986). Typically, the development of reasoning ability is associated with individuals' metacognition, interpersonal communication, and developmental growth. Having the ability to reason is a part of an individual's ability to performing mental operation, which can be affected by external learning styles and internal cognitive styles. A cognitive style describes how an individual perceives, remembers, thinks, and solves problems in different contexts (Lomberg, Kollmann, & Stöckmann, 2017; Volkova & Rusalov, 2016) and it varies by person. Therefore, cultivating reasoning ability has to account for cognitive styles. Learners with different cognitive styles may need different learning strategies and processes to facilitate effective learning.

Riding and Sadler-Smith (1997) pointed out that learning strategies designed specifically for individuals with different cognitive styles have a better chance of increasing the efficiency and effectiveness of learning and assisting learners with overcome learning difficulties. It is important to include cognitive styles when one studies pedagogical strategies. Otherwise, some strategies that favor one group could result in no effect for the other. To use digital games for learning effectively, it is important to examine whether and how playing digital games affect students with different cognitive styles. Research has shown that students with different cognitive styles demonstrate different preferences in learning and social adaptation (Chen & Chang, 2016). The authors found that if students have a cognitive style that is similar to that of their teacher, they have a higher chance of reporting a more positive learning experience. This study shows that the effect of learning for students with different cognitive styles varies.

Recent research in learning technology attempted to exploit digital games to help students learn reasoning and problem-solving skills. Young children play games to build their self-esteem and self-efficacy, to acquire metacognition and motor skills, to practice interpersonal and social communication, to improve developmental growth, to participate in role play, and to exercise emotional expression (Broadhead, 2006; Erhel & Jamet, 2013; Kennewell & Morgan, 2006; Li & Tsai, 2013; Moreno, 2012). These studies showed that playing digital games

could be more than solely for entertainment; it can also be an effective approach to cultivate children's reasoning ability. Salient issues such as relationships among digital games, reasoning ability, and cognitive style, however, remain unaddressed, leaving ample opportunities for further investigations.

This study aims to investigate whether a child's reasoning ability can be facilitated in a digital adventure gaming environment and whether cognitive styles—field-independent (FI) vs. field-dependent (FD)—affect the acquisition of reasoning ability in the gaming environment. In short, FI and FD are distinguished by the ability to discern detail information from its surrounding environment, where people with FD style being relatively weaker than those with FI. Detailed descriptions of these two cognitive styles are provided in the next section (Literature Review). We expect to see a positive learning experience of using digital games in promoting children's reasoning ability. Additionally, we expect that their cognitive styles will affect the outcome of using digital games to facilitate the acquisition of reasoning ability. Four research questions were studied as follows:

- Does a child's achievement score on a pretest and a posttest show significant differences after experiencing digital game playing?
- Does the achievement score of children with different cognitive styles differ after experiencing digital game training and paper-based training?
- Does a child's pattern of game playing show reliable difference according to their cognitive style?
- Will a FI child tend to think more independently and not require much external assistance than a FD child?

## **2. Literature review**

It is known to researchers that field-dependent and field-independent learners prefer different instructional models and materials. It is not clear how these learners would behave in and learn from playing video games. The correlation between video game playing and reasoning ability is also unclear. In this section, we review some key concepts and recent studies related to reasoning ability, video games, and cognitive styles.

### **2.1. Reasoning ability and its development**

Reasoning ability normally described as one of higher-order thinking skills (Krulik & Rudnick, 1993) and is an essential ability when dealing with real-world problems. It allows an individual to use prior knowledge with new information and apply principles systematically to construct the relation between old and new problems (Rosser, 1994; Spitz, 1979). This mental process enables a person to make logical arguments (Barbey & Barsalou, 2009) apply logical rules (Wilhelm, 2005) and understand casual relations in an environment (Piaget & Inhelder, 2008). Because learning cannot possibly cover all known situations, reasoning ability becomes especially crucial for preparing an individual for future unknown problems. We regard reasoning ability as an individual's ability to deliberately use known information to solve an unseen problem.

The development of cognitive skills is a gradual process, from simple and concrete to complex and abstract. The participants in the study were 6<sup>th</sup> graders who, according to Piaget's stages of cognitive development, were between the late concrete operational stage and the early formal operational stage. In the concrete optional stage, students tend to have logical reasoning thinking of concrete issues and concepts of classification and sequence. In the formal operational stage, students develop the ability to think about abstract concepts and are able to think logically—a systematic and logical process (Piaget & Inhelder, 2008). This is a crucial period when students transition from concrete to abstract reasoning; therefore, they should be provided with appropriate teaching aids. Further, developmental training is important for students during this period. Various challenges and puzzles in a digital adventure game were used to train the participants of this study. In general, three reasoning methods—deductive reasoning, inductive reasoning, and analogical reasoning—are used in problem solving. Deductive reasoning is used to verify hypotheses, inductive reasoning is used to formulate general rules, and analogical reasoning is used to apply general rules to similar situations. In this study, participants needed to use available information from the game scenarios, combine it with prior knowledge as the basis to perform deductive, inductive, or analogical reasoning to identify and solve problems by manipulating rules—a reasoning process (Wilhelm, 2005). We believe it is plausible to use digital adventure games to cultivate reasoning ability.

## 2.2. Reasoning ability and digital games

Previous research studied the effects of digital games on reasoning ability and found that these games could be used effectively to enhance children's reasoning abilities (Bakker et al., 2015; Bottino et al., 2007; Liu & Lin, 2009). Because problem solving abilities intertwine with reasoning abilities, we included studies about digital games and problem solving in this section as well.

Mather (1986) reported that adventure games could improve students' reading skills, cultivate their creativity, and enhance their problem-solving ability. Australian researchers conducted a study on elementary students to find whether playing adventure games helped students' learning, and the results showed that these games could in fact improve students' problem-solving ability and skills (Grundy, 1991). Dempsey, Lucassen, Haynes, and Casey (1996) conducted a study on 40 adults with regards to adventure games and learning. Their results illustrated that these games were beneficial for problem-solving and decision-making abilities. In addition, Amory, Naicker, Vincent, and Adams (1998) conducted a study on 20 college underclassmen in England to find the most applicable educational digital games and interesting or helpful game elements. The results showed that adventure games could combine pictures, sounds, and stories to improve students' logic, memory, imagination, and problem-solving ability. Hsiao et al. (2014) discussed how adventure games affected 5<sup>th</sup> graders' creativity, problem-solving ability and achievement motivation. The results indicated that the experimental group had higher scores on the posttest of the problem-solving assessment than the control group did. These studies suggested that playing adventure games could develop general problem-solving abilities.

The results of the previous mentioned studies indicated that digital games positively affected problem-solving and that reasoning ability was associated with problem solving. Reasoning abilities and problem-solving abilities are often considered complementary to each other (Jenny & Claire, 2008; Krulik & Rudnick, 1993). Reasoning out the answer requires a student to examine if the solution is logical and plausible. Aside from chance or luck, students must know how and where to find the solution to a problem. Puzzle games emphasize on solving problems and often require the players to use reasoning ability with given information and available objects in a novel situation. Therefore, puzzle adventure games should have positive effects on problem solving, and the study focused on whether puzzle adventure puzzle games affected reasoning ability positively (Bakker et al., 2015; Crompton et al., 2018). Players' acceptance and adaptability of digital games vary slightly due to different cognitive styles, methods of processing information, individual cognitive capacity, thinking ability, and abilities to generalize symbols (Lin et al., 2011). In this study, we analyzed how playing a puzzle adventure game affected players' reasoning ability and discuss how players with different cognitive styles approach problems and obstacles when playing this puzzle adventure game.

Digital adventure games are video games that players control characters to interact with objects or other computer-generated characters to solve problems or puzzles in an artificially created digital world (Cavallari, Hedberg, & Harper, 1992). They normally contain adventure stories with rich context. In such a simulated world, players can try many actions, such as opening a door, throwing a rock, combining two objects to solve different problems or challenges. They often need to decode messages, make hypotheses, or apply inferences in their journey of the story (Chandler & Chandler, 2011). Regarding the benefit of playing digital adventure games, Ju and Wagner stated that reasoning and problem-solving skills are required in adventure games (Ju & Wagner, 1997).

## 2.3. Cognitive styles

Researchers have attempted to measure different cognitive styles and identify characteristics of different dimensions of cognitive style so that they could better understand human mental operation. Messick (1984) proposed approximately 20 cognitive styles and there were over 30 cognitive styles proposed (Riding & Cheema, 1991). Nevertheless, some of them were repetitive or excessively overlapping with one another. Riding and Cheema (1991) proposed two main orthogonal cognitive style families, "wholistic-analytic" and "verbal-imagery", based on their review of cognitive style. Previous studies on teaching and learning mainly focused on analyzing learning achievement and attitude with regard to field-dependent (FD) and field-independent (FI) (Sadler-Smith, 2001; Riding & Rayner, 2013). Because "FD vs. FI" was studied most widely and many assessment tools for learner's performance in digital games have been validated, we adapted the "FI vs. FD" paradigm in this study.

Based on their research results, Witkin, Moore, Goodenough, and Cox (1977) identified that there are various differences between FD and FI people. FI people had a tendency to be more autonomous in relation to the development of cognitive skills and less autonomous in relation to the development on interpersonal skills;

conversely, FD people had a tendency to be more autonomous in relation to the development of high interpersonal skills and less autonomous in relation to the development of cognitive restructuring skills. In addition, FI people preferred individualized learning whereas FD ones enjoyed cooperative learning. Studies (Chen & Chang, 2016; Lomborg et al., 2017; Lugli et al., 2017) indicated that FD students tended to emphasize a certain aspect and searched for solutions based on certain casual relations or reasons. These students also tended to rely on external cues to observe subjects, make a judgment, and needed constructed materials to learn knowledge. FI students, on the other hand, tended to organize learning materials based on the understood casual relations and analyses of reasons. Although their cognitive styles were different, FD and FI students' intelligence and intellectual level were not directly correlated. Students with different cognitive styles might have the same intelligence or intellectual level (Tamaoka, 1985). They simply thrived under different learning conditions. Two studies showed that different cognitive styles affected students' learning behaviors, and those with FD and FI appeared to have different academic performances due to pedagogical strategy (Chang, Lin, & Chen, 2019; Chen & Macredie, 2002). Researchers suggested that teacher must adapt their instruction to students with different cognitive styles and provide necessary assistance to achieve a better learning outcome (Mefoh et al., 2017; Thomas & McKay, 2010).

Researchers developed Embedded Figures Test (EFT) to categorize an individual as FD or FI. According to the EFT, those who tend to rely on external cues and are less able to differentiate an embedded figure from an organized field are labeled as FD while those who tend to rely on internal reasoning and are better at differentiating an embedded figure from an organized field are labeled as FI. Therefore, FI people are able to analyze a larger complex figure, distinguish discontinuous parts from the figure by ignoring irrelevant information, and extract the embedded figures, and coordinate the embedded figures as obligatory in the organized field; FD people tend to view the organized field as a whole and thus are unable to eliminate unrelated parts of the complex figure. Therefore, how students with FD or FI perform in a puzzle adventure game with rich pictures and different information is worthy of investigation.

Parkinson and Redmond (2002) investigated the relations among cognitive styles, learning outcomes, and three different types of learning media: texts, multimedia CD-ROM, and the Internet. Lee et al. (2005) explored the relations between cognitive styles and learning preferences in a fundamental multimedia course of the hypermedia learning system. Mampadi, Chen, Ghinea, and Chen (2011) discussed the differences of the students with FD and FI cognitive styles in the linear and non-linear learning during the digital game play. However, these studies did not analyze the design and content of the digital game.

Other studies showed that different cognitive styles affected students' learning behaviors, and those with FD and FI cognitive styles had different academic performances (Chen & Macredie, 2002; Salih & Erdat, 2007). Teachers needed to adapt their teaching instruction to students with different cognitive styles and learning methods and provide necessary assistance to achieve a better learning outcome (Chen & Macredie, 2002; Hansen, 1997; Riding & Sadler-Smith, 1997).

### **3. Research method and procedure**

To study the effect of playing video games on reasoning ability and whether students with different cognitive styles benefit from playing puzzle adventure games equally, we designed our experimental study with three groups. One group played a puzzle adventure game, one group was trained by solving reasoning problems on papers, and one group did not receive any treatment. In this study, participants in group one needed to use available information from the game scenarios, combine it with prior knowledge as the basis to do deductive, inductive, or analogical reasoning identify and solve problems by manipulating rules to, which in essence is a reasoning process (Wilhelm, 2005). We believe playing puzzle adventure games has the potential of cultivating reasoning ability. We detail our study procedure this section.

#### **3.1. Intervention instruments**

The digital game group (T1) played a puzzle adventure game called *Machinarium* (Figure 1) in the experiment. In this game, players control a character by using a mouse to point-and-click to solve a series of puzzles and brain teasers that require reasoning ability. The game contains no human language conversation, so players have to rely on observing objects on the scene, making inference of their relations, and connecting related ones to help them solve the puzzles. The game includes five levels. There are several challenges in each level and several puzzles in each challenge in the game. To solve the puzzles, players must apply problem-solving skills, which

cultivates reasoning ability and promotes higher order thinking skills. While participants were playing the game, their mouse clicks and movements were recorded using a program called Morae Recorder. The recorded logs allow us to analyze players' behavior patterns.



*Figure 1. A snapshot of the puzzle adventure game cover*

Another group of participants received a paper-based training (T2) with the same amount of time and frequency as participants in T1. The training is based on a book that trains logic thinking and reasoning for children around age 12 or above. The training on the paper is text-based descriptions and activities. The no-treatment group (T3) received neither the video game nor the paper-based training during that time.

### **3.2. Assessment of reasoning abilities**

Raven (1936) developed the Ravens's Progressive Matrices (RPM) to measure the reasoning component of Spearman's *g*, which consists of the two principles in cognition, education of relations (i.e., induction and analog ability) and education of correlations (i.e., interpretation ability). Standard Progressive Matrices (SPM) was the original version of the family of the matrices. Other matrices such as the Colored Progressive Matrices Parallel (CPM-P), the Standard Progressive Matrices Parallel (SPM-P), and the Standard Progressive Matrices Plus (SPM<sup>+</sup>) were published after SPM. We chose SPM-P because it is designed for children of age between 10 to 12 and it has high reliability. The SPM-P consists of 60 items that are evenly divided into five series. Each item consists of an incomplete pattern (matrix) that the subject is to find the matching piece out of six choices shown beneath the matrix. The internal consistency reliability of the SPM-P is between 0.83 to 0.90; the split-half reliability is between 0.87 to 0.92; the test-retest reliability of the five-week study is 0.81.

### **3.3. Experimental procedure**

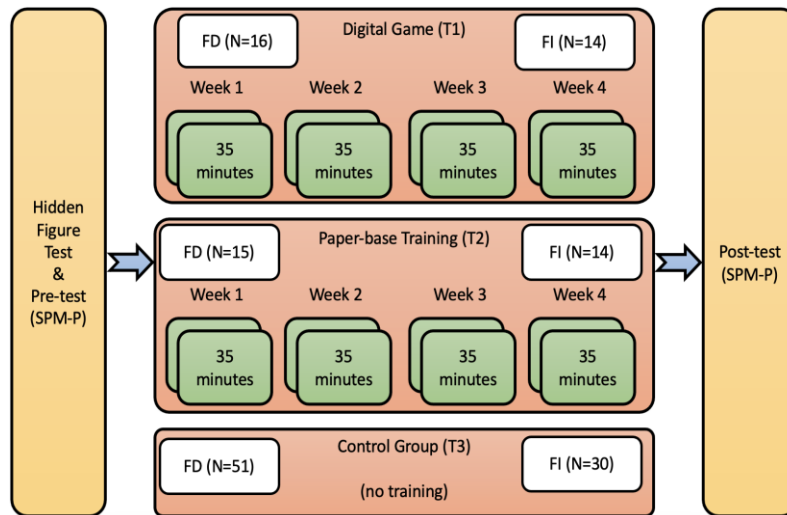
The participants of this study were 140 6<sup>th</sup> graders at an elementary school in Southern Taiwan. They were randomly assigned into three groups, treatment group 1 (T1), treatment group 2 (T2), and a control group (T3), based on the scores of a cognitive style inventory called Hidden Figure Test (HFT). According to the HFT test score, participants were divided into three groups: FI, FD, and indeterminate. In the FI group, participants were randomly assigned to T1, T2, and T3. The same assignment was done for the FD group. Finally, the participants in the indeterminate group were randomly assigned to T3. The diagram in Figure 2 shows the details of our experimental procedure. The participants whose scores on the HFT are higher than the mean of the highest score and the lowest score are categorized as FI participants. Those whose scores on the HFT lower than the mean of the highest score and the lowest score are categorized as FD participants. Some participants were moved to the control group in order to afford each participant a personal computer in T1 and to balance the number of FI and FD participants in T1 and T2. Table 1 shows the final number of participants in each group.

The participants received a pretest prior to the experiment and a posttest after the experiment. Those in the T1 played the puzzle adventure game for eight times in four weeks, each lasted 35 minutes. Students in the T2

received paper-based training for eight times, again, each lasted 35 minutes. The T3 was regarded as the control group, which received no training activities in this study.

*Table 1. Distribution of participants*

Participants	T1 (Puzzle Adventure Game)	T2 (Paper-based Training)	T3 (Control)
FD	16	15	51
FI	14	14	30
Total	30	29	81



*Figure 2. Experimental procedure diagram*

### 3.4. Research design

The research design in the study was a randomized pretest-posttest with a control group design. The two independent variables were game playing and paper-based training. The dependent variables were the SPM-P criterion tests that were given immediately after the participants finished the treatment.

A Multivariate Analysis of Variance (MANOVA) was used to analyze data. The main effects and the potential interaction of the two independent variables were examined. Where significant *F*-values were found, pair-wise multiple comparison tests were performed using the Scheffe test.

## 4. Findings

### 4.1. Results of multivariate analysis of variance (MANOVA)

This research design of the study is a randomized 2 x 3 pretest and posttest design. The two independent variables are: cognitive styles (FD and FI) and treatment types (T1: puzzle adventure game; T2: paper-based training; T3: no training). The dependent variable was students' reasoning ability measured by the Standard Progressive Matrices Parallel (SPM-P). A multivariate analysis of variance (MANOVA) was conducted to analyze the collected data from 140 participants.

A descriptive statistics summary including both pretest and posttest is illustrated in Table 2.

*Table 2. Means and standard deviations of the SPM-P measurement*

Test	FI						FD					
	T1 ( <i>n</i> =16)		T2 ( <i>n</i> =15)		T3 ( <i>n</i> =50)		T1 ( <i>n</i> =14)		T2 ( <i>n</i> =14)		T3 ( <i>n</i> =31)	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Pretest	106.57	14.92	94.07	17.84	96.25	17.19	104.50	13.92	104.54	16.50	95.22	15.49
Posttest	118.07	10.24	103.87	12.99	99.83	18.38	108.23	15.68	104.00	18.49	96.45	16.68

The results from the MANOVA (shown in Table 3) indicated that the two independent variables (cognitive style vs. training) had no statistically significant interaction (for the pretest,  $F = .15$ ;  $p = .864$ ; for the posttest,  $F = .75$ ;  $p = .477$ ). Therefore, it is valid to analyze the effects of cognitive styles and gaming on student's reasoning ability independently.

Table 3. MANOVA analysis of between-subjects effects

Source	Dependent Variables	df	SS	MS	$F$	$p$
Cognitive Styles	Pretest	1	2575.27	2575.27	10.55	.001*
	Posttest	1	2588.30	2588.30	10.25	.002*
Experimental Groups	Pretest	2	7.82	3.91	.02	.984
	Posttest	2	1718.11	859.05	3.40	.036*
Cognitive Styles $\times$ Experimental Groups	Pretest	2	72.25	36.13	.15	.864
	Posttest	2	376.638	188.32	.75	.477
Error	Pretest	129	31477.97	244.02		
	Posttest	129	32590.32	252.64		
Total	Pretest	135	1366422.00			
	Posttest	135	1475671.00			

Note. \* $p < .05$ .

Table 2 and Table 3 also showed that regardless of experimental treatments, FI students performed reliably better than the FD students in both pretest ( $F = 10.55$ ,  $p = .001$ ) and posttest ( $F = 10.25$ ,  $p = .002$ ). Among the experimental groups, a significant difference was found in students' posttest ( $F = 3.40$ ,  $p = .036$ ). A follow-up Scheffe multiple comparison test was conducted to discern the significant difference (illustrated in Table 4). We found that the difference came from the gaming group and the control group ( $p = .018$ ).

Table 4. Scheffe multiple comparison of groups

Source		Mean difference	Std. Err.	Significance
Pretest	T1 & T2	-.456	4.263	.994
	T1 & T3	1.449	3.380	.912
	T2 & T3	1.905	3.573	.868
Posttest	T1 & T2	8.724	4.337	.137
	T1 & T3	9.909*	3.439	.018*
	T2 & T3	1.185	3.636	.948

Note. \* $p < .05$ .

## 4.2. Results of univariate analysis of variance (ANOVA)

According to Table 5, a significant difference was found in the posttest between FI and FD students ( $F = 10.59$ ,  $p = .003$ ) in T1 (gaming). By considering the results in Table 2, we found that FI students (mean = 118.07; standard deviation = 10.24) obtained a significant better reasoning ability than the FD students (mean = 103.87; standard deviation = 12.99) in the gaming group.

Table 5. ANOVA analysis of posttest for Treatment 1

Source	Dependent variables	DF	SS	MS	$F$	$p$
Pretest	Between groups	1	1132.39	1132.39	4.16	.051
	Within groups	27	7348.36	272.162		
	Total	28	8489.69			
Posttest	Between groups	1	1461.13	1461.13	10.59	.003*
	Within groups	27	3724.66	137.95		
	Total	28	5185.79			

Note. \* $p < .05$ .

In T2 (paper-based training), no significant difference was found in both pretest ( $F = 1.513$ ,  $p = .231$ ), and posttest ( $F = 0.319$ ,  $p = .578$ ) between FI and FD students, according to Table 6.



Table 6. ANOVA analysis of posttest for Treatment 2

Source	Dependent variables	DF	SS	MS	<i>F</i>	<i>p</i>
Pretest	Between groups	1	428.68	428.68	1.513	.231
	Within groups	23	6517.48	283.369		
	Total	24	6946.16			
Posttest	Between groups	1	108.33	108.33	.319	.578
	Within groups	23	7821.67	340.07		
	Total	24	7930.00			

Note. \* $p < .05$ .

In T3 (control), significant differences were found in both pretest ( $F = 7.30$ ,  $p = .008$ ) and posttest ( $F = 9.84$ ,  $p = .002$ ) between FI and FD students. By considering the results in Table 7, we found that in the control group, FI students performed reliably better in reasoning ability test than the FD students in both pretest and posttest.

Table 7. ANOVA analysis of posttest for Treatment 3

Source	Dependent variables	DF	SS	MS	<i>F</i>	<i>p</i>
Pretest	Between groups	1	1628.19	1628.19	7.30	.008*
	Within groups	79	17612.18	222.94		
	Total	80	19240.32			
Posttest	Between groups	1	2622.29	2622.29	9.84	.002*
	Within groups	79	21043.99	266.38		
	Total	80	23666.22			

Note. \* $p < .05$ .

#### 4.3. Pattern of mouse clicks

We used Morae Manage to analyze the recordings of the participants' game playing behavior and used the mouse clicks search options to search for all mouse clicks that occurred during a particular time span (i.e., 10 seconds). Figure 3 shows the mouse movements of participants with two cognitive styles in 10 seconds while solving two challenges of level one. The left column illustrates the FI participants' mouse movements while the right column illustrates the FD participants' mouse movements. Players have to search for a doll in the first challenge. As Figure 3 has shown, the traces for FI participants (Figure 3(a)) are more condensed than those on the right (Figure 3(b)). This suggests that the FI participants carefully observed the details around the robot, and the FD participants moved the mouse pointer all around the screen to search for the doll. Similar patterns are observed in another challenge of level one, which are shown at the bottom row of Figure 3.

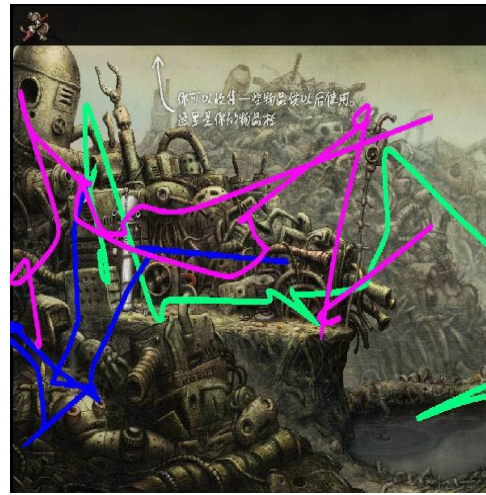
Because there is a short movie that serves as a hint about the goals of upcoming challenge only in the beginning of level one and level two, we also provide an example pattern from level three when the short movie was not shown to the participants. As suggested in Figure 4, the FI carefully observed the details of the screen. On the other hand, the FD participants moved their mouse pointer all around the screen and clicked the mouse button many times. The pattern is similar to what Figure 3 shows. It illustrates that the FI participants tended to think and analyze the relations and that the FD participants were weaker in reasoning and analytical skills, with or without hints.

The number of times that FI and FD participants clicked a mouse button in three challenges (1<sup>st</sup>, 2<sup>nd</sup>, and 5<sup>th</sup>) of level three is illustrated in Figure 5. The blue diamonds represent data points for the FI participants, and the red diamonds represent data points for the FD participants.

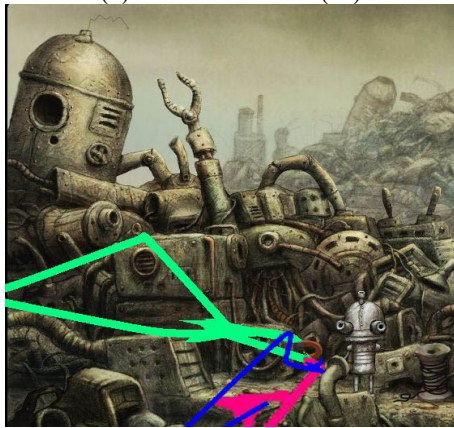
The results show that the FD participants used much more mouse clicks than the FI participants. A possible explanation is that FI participants tended to think and analyze problems while playing puzzle adventure games and clicking the mouse button was not intentional. However, the FD participants might not carefully analyze the tasks to be completed, which caused quite a few unnecessary random clicks.



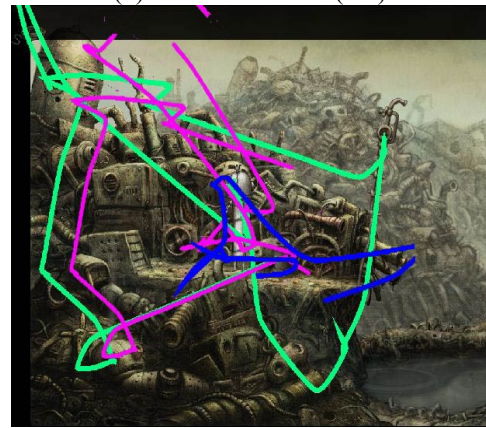
(a) Search for a doll (FI)



(b) Search for a doll (FD)



(c) Search for a magnet and a line (FI)

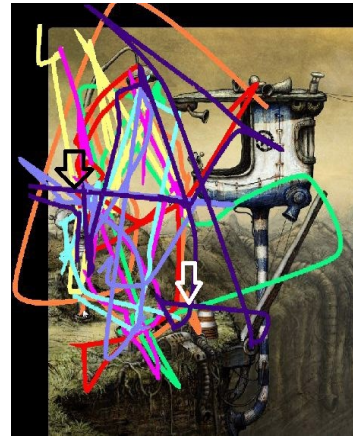


(d) Search for a magnet and a line (FD)

Figure 3. Examples of traces of mouse movements from FI and FD participants in two challenges of level one

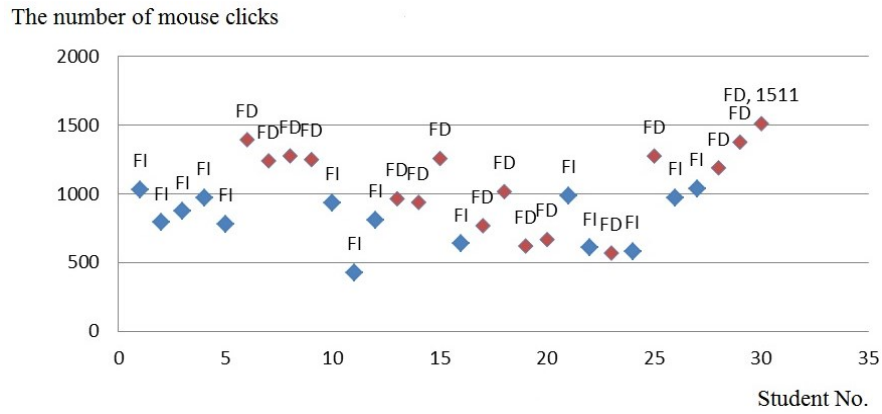


(a) Light up the lamp (FI)

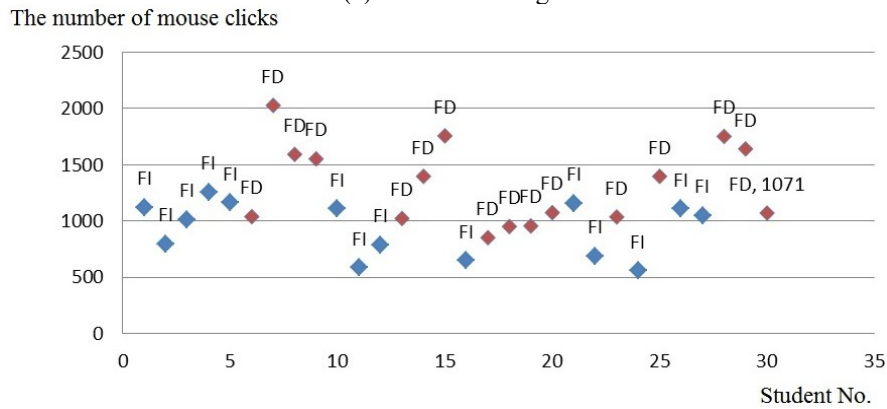


(b) Light up the lamp (FD)

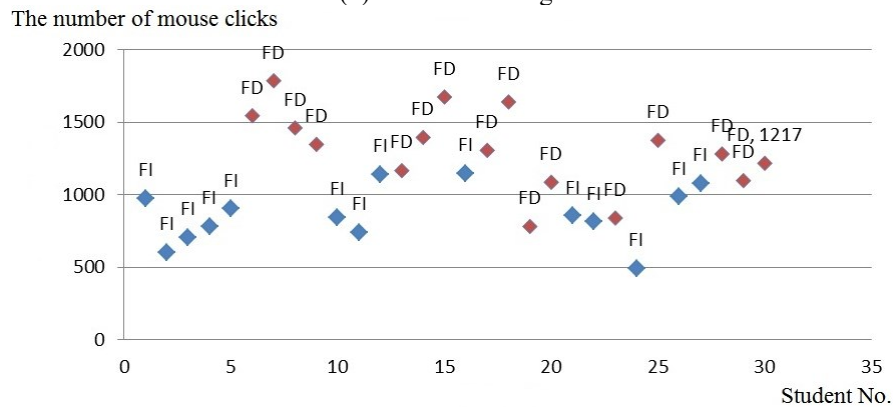
Figure 4. Examples of traces of mouse movements for FI and FD participants on level three



(a) The 1<sup>st</sup> challenge



(b) The 2<sup>nd</sup> challenge



(c) The 5<sup>th</sup> challenge

Figure 5. Numbers of mouse clicks on level three

## 5. Discussion

From the results, we see that playing the puzzle adventure game improves students' reasoning ability. Further looking at the behavior patterns suggests that students with different cognitive styles use different information seeking strategies. We discuss our results in detail below along with suggestions for further research.

### 5.1. Effects of puzzle adventure games on reasoning ability

For research question one, does a child's achievement score on a pretest and a posttest show significant differences after experiencing puzzle adventure game playing, we found that participants' score did increase reliably after playing the puzzle adventure game. According to SPM-P scores, there are significant differences in student's reasoning ability among three groups—puzzle adventure game, paper-based training, and no training (control group). Regardless of student's cognitive style, the puzzle adventure game group perform better than those in the control group on the posttest. This finding is consistent with Liu and Lin's results (2009) that

indicated playing digital puzzle games improved the players' reasoning ability. The items of SPM-P are non-verbal, multiple-choice questions that test takers must identify the missing elements from a given pattern. Similarly, the puzzle adventure game requires players to solve a series of puzzles using their reasoning ability with no verbal information. Because both the SPM-P test and the puzzle adventure game require the participants to observe the given clues and the surrounding details to solve puzzles using their reasoning ability, it is not surprise that the puzzle adventure game group outperform other groups.

On the other hand, there is no significant difference on the posttest between the paper-based training group and the control group. A possible explanation is that the paper-based training activities require the participants to read, understand, and analyze verbal information to find out the answers. This type of training may be beneficial for verbal reasoning but not for graphical reasoning. As mentioned previously, SPM-P may favor non-verbal reasoning ability, which may not benefit verbal training such as what we did in the paper-based training group. Another reason could be the lack of motivation to learn and to solve the problems. This verbal type of reasoning that paper-based training provides may not attract the participants' attention to solve problems.

With regard to research question two, does the achievement score of children with different cognitive styles differ after experiencing puzzle adventure game training and paper-based training, an interaction between cognitive styles and puzzle adventure gaming effects is also identified. Within the puzzle adventure game group (T1), the FI participants had a reliably higher posttest scores than the FD participants ( $F = 10.59, p = .003$ ). This means that after the treatment, regardless of training types, FI learners perform reliably better than FD learners in the posttest measurement. This finding is consistent with previous work that studied cognitive styles and student learning outcome. Based on their research results, Witkin et al. (1977) indicated that FD and FI people have great differences in several ways. For example, FI people have a tendency to be more autonomous in relation to the development of cognitive restructuring skills and less autonomous in relation to the development on interpersonal skills; conversely, FD people have a tendency to be more autonomous in relation to the development of high interpersonal skills and less autonomous in relation to the development of cognitive restructuring skills. In addition, the FI people enjoy individualized learning while the FD ones enjoy cooperative learning. For the paper-based training group (T2), we did not find reliable differences between FI and FD in posttest after treatment. This indicates that there is no difference in posttest between FI and FD in using paper-based training.

## 5.2. Effects of cognitive styles on reasoning ability

The analysis of the traces of mouse movements, which answers research question three (does a child's pattern of puzzle adventure game playing show reliable difference according to their cognitive style?), indicates that the FI participants tend to observe surrounding details, to search for appropriate prompts from a small to a large area gradually, and to consider possible solutions when solving problems. We believe that they demonstrate independent thinking and that they tend to apply analytical skills without relying on a lot of clues. On the other hand, the FD participants tend to search for appropriate prompts from a large to a small area and think less carefully about solutions. They tend to rely on the given clues in a form of a thought bubble and external assistance from the teacher or the other participants.

*Table 8. Differences between FI and FD participants*

ID	FI Participants	FD Participants
D1	Carefully thinking before taking actions	Expecting to have more external information
D2	Paying less attention to text descriptions	Paying more attention to text descriptions
D3	Detailed observation of the field	Extensively observing of the field
D4	Less clicking on text descriptions	Often clicking on text descriptions
D5	Taking longer time to combine prompts because of paying less attention to text explanations	Taking shorter time to combine prompts because of paying more attention to text explanations
D6	Considering when and where to use new prompts when receiving them	Storing new prompts until encountering difficulties
D7	Less interaction and discussion with others	More interaction and discussion with others

In addition, the analysis of the numbers of mouse clicks, which answers research question four (will a FI child tend to think more independently and not require much external assistance than a FD child?), indicates that the FI participants tend to independently think before taking actions whereas the FD participants tend to rely on external assistance to solve the puzzles or attempted to ask for useful information. This result is consistent with previous findings from the review of literature, which indicated that the FI individuals generally were analytical in their approach to solve problems, whereas the FD individuals were more global in their approaches and tended

to rely on external assistance to solve problems (Mampadi et al., 2011; Salih & Erdat, 2007). Table 8 summarizes our observations about the differences of the FI and FD participants while playing the puzzle adventure game.

Kozhevnikov (2007) found that the FI students tended to concentrate on a certain aspect and searched for solutions based on certain casual relations or reasons. FD students usually relied on external cues when observing subjects and making judgments, and they needed more help for scaffolding. The FI students tended to organize learning materials based on the known casual relations and on analyses of reasons. Students' intelligence and intellectual level were not directly correlated to either FI or FD.

While playing the puzzle adventure game, the FD learners tend to process information from a global perspective; they are relatively unable to construct their knowledge under a non-structured environment, so they tend to learn passively and rely on external assistance from teachers or classmates. This conclusion is supported by our observation that FD participants tend to click on more prompts than the FI participants did. However, it seems the text descriptions in the prompts of the puzzle adventure game are not sufficient to support the FD participants to solve designated tasks. The future design for the puzzle adventure game could revise its prompt system by considering participants with different cognitive styles or it can be adaptive. For example, the information that the player receives can change based on the number of times the prompt is clicked. To discourage abusing the prompt feature, the points received could decrease as the degree of details increases.

Unlike the FD players, the FI players in the study are analytical and able to construct their knowledge independently so that they do not need the prompt system as much. The linear storyline of the puzzle adventure games may be appropriate for the FI individuals. Non-linear storyline of the puzzle adventure game may be better for the FD players because it did not require players to finish the current level to advance to the next level. The FD players need more text description in the prompts than the FI players to complete a level. On the other hand, FI individuals need more figural questions to improve their reasoning ability.

### 5.3. Suggestions for further research and puzzle adventure game design

From our observations and analyses, FD and FI participants illustrate some distinct behaviors and preferences (summarized in Table 8). Accordingly, we propose some suggestions for designers of puzzle adventure games to target players of different cognitive styles (see Table 9). Puzzle adventure games could offer them as options to the players and allow them to select the ones they prefer (without preferring the options to FI or FD). This approach could potentially maximize the learning benefit and enjoyment. Because there are endless possibilities for the develop the storylines of a puzzle adventure game, these guidelines are intended for the interactivity and interface design and not for the game content. The labels in the first column (reasons) correspond to the differences in Table 8. The second and third columns describe our suggestions for each type of players.

*Table 9. Design suggestions for FI and FD players based on the differences in Table 8*

Reasons	FI Players	FD Players
D1, D5	Fewer prompts that contain simple texts	More prompts in a level The more times to click on prompts, fewer points players receive when completing a level
D1, D2, D5	Fewer text descriptions in a level	More text descriptions in a level
D1, D2, D5	More figural questions for reasoning and thinking	More text descriptions of missions
D2, D5	Basic text descriptions of prompts	More text descriptions of prompts
D3	Linear storyline of the puzzle adventure game that requires players to finish the current level and then proceed to the next level	Non-linear storyline of the puzzle adventure game that doesn't require players to finish the current level to proceed to the next level
D3, D4, D6	No hidden prompts	Provide hidden prompts to facilitate players to complete the level

Table 9 provides guidelines for future puzzle adventure game design for FI and FD players. In this table, we list suggestions based on our observation on the differences (illustrated in Table 8) between FI and FD players. According to our findings, future research should continue to investigate the impact of digital gaming environments on students' learning achievement, especially on their higher-order reasoning ability such as problem-solving and critical-thinking skills. In addition, future studies should continue to investigate other human factors in a digital gaming environment such as learners' individual differences, learning styles, and

preferences in using visual/audio materials. Many of the independent variables associated with the study of aptitude-treatment interactions should be taken into account in the design of digital gaming environment.

While digital gaming environment may be manipulated to positively influence students' reasoning ability, special attention must be given to concrete game design guidelines derived from reliable experimental methodologies, as well as to consideration of learner characteristics and styles. Only by conducting a systematic investigation where learning variables are judiciously manipulated to determine their relative effectiveness and efficiency of facilitating specifically designated learning objectives will the true potential inherent in digital game design be realized.

## 6. Conclusion

Reasoning ability is an important cognitive skill for solving real-world problems and puzzle adventure games provide an enjoyable and engaging environment where players can experience different reasoning skills. In addition, it is known that people have different cognitive styles, so it is expected that people benefit unevenly in the same learning environment. To study the effect of puzzle adventure games on reasoning ability for players with different cognitive styles, we studied elementary students who are in the process of developing reasoning ability. We compared their pretest and posttest scores on reasoning ability, measured by SPM-P. We discussed several findings from our data. First, students in the puzzle adventure game group score reliably higher in the posttest than those who do not play the game. Second, FI participants benefit more than the FD participants with regard to improvement in reasoning ability after playing the puzzle adventure game. Lastly, FI and FD participants show different playing behavior patterns (i.e., global vs. linear).

Based on our findings, playing puzzle adventure games helps elementary school children improve their reasoning ability, especially for those who are FI. We feel it may be the case that these games are engaging, and students are able to interact with the game scenarios to see the outcomes of their actions immediately. Our work was designed to provide additional empirical evidence in game-based learning and expand the effect of game-based learning to learner factors (cognitive styles.) As mentioned in the future study, research studies can build on our results and techniques to deepen our understanding of game-based learning in action.

## Acknowledgement

The research reported in this paper has been supported in part by the Ministry of Science and Technology in Taiwan under the research project number MOST 108-2511-H-024-009, MOST 108-2918-I-024-002 and MOST 109-2511-H-024-002. We would like to thank Catherine Yeh for her assistance on improving the paper. The anonymous reviewers are appreciated for their valuable comments.

## References

- Amory, A., Naicker, K., Vincent, J., & Adams, C. (1998, June). *Computer games as a learning resource*. Paper presented at the Proceedings of ED-MEDIA, South Africa.
- Bakker, M., van den Heuvel-Panhuizen, M., & Robitzsch, A. (2015). Effects of playing mathematics computer games on primary school students' multiplicative reasoning ability. *Contemporary Educational Psychology*, 40, 55-71.
- Barbey, A. K., & Barsalou, L. W. (2009). Reasoning and problem solving: Models. *Encyclopedia of Neuroscience*, 8, 35-43.
- Bottino, R. M., Ferlino, L., Ott, M., & Tavella, M. (2007). Developing strategic and reasoning abilities with computer games at primary school level. *Computers & Education*, 49(4), 1272-1286.
- Broadhead, P. (2006). Developing an understanding of young children's learning through play: The place of observation, interaction and reflection. *British Educational Research Journal*, 32(2), 191-207.
- Carroll, J. B. (1993). *Human cognitive abilities: A Survey of factor-analytic studies*. Cambridge, England: Cambridge University Press.
- Cavallari, B., Heldberg, J., & Harper, B. (1992). Adventure games in education: A Review. *Australasian journal of educational technology*, 8(2), 172-184.
- Chandler, H. M., & Chandler, R. (2011). *Fundamentals of game development*. Sudbury, MA: Jones & Bartlett Learning.



- Chang, J. J., Lin, W. S., & Chen, H. R. (2019). How attention level and cognitive style affect learning in a MOOC environment? Based on the perspective of brainwave analysis. *Computers in Human Behavior*, 100, 209-217.
- Chen, S. Y., & Chang, L. P. (2016). The Influences of cognitive styles on individual learning and collaborative learning. *Innovations in Education and Teaching International*, 53(4), 458-471.
- Chen, S. Y., & Macredie, R. D. (2002). Cognitive styles and hypermedia navigation: Development of a learning model. *Journal of the American society for information science and technology*, 53(1), 3-15.
- Crompton, H., Lin, Y. C., Burke, D., & Block, A. (2018). Mobile digital games as an educational tool in K-12 schools. In *Mobile and Ubiquitous Learning* (pp. 3-17). Springer, Singapore.
- Dempsey, J. V., Lucassen, B. A., Haynes, L. L., & Casey, S. M. (1996, April). *Instructional applications of computer games*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.
- Erhel, S., & Jamet, E. (2013). Digital game-based learning: Impact of instructions and feedback on motivation and learning effectiveness. *Computers & Education*, 67, 156-167.
- Grundy, S. (1991). A Computer adventure as a worthwhile educational experience. *Interchange*, 22(4), 41-55.
- Hansen, J. W. (1997). Cognitive styles and technology-based education. *Journal of Technology Studies*, 23(1), 14-23.
- Hsiao, H. S., Chang, C. S., Lin, C. Y., & Hu, P. M. (2014). Development of children's creativity and manual skills within digital game-based learning environment. *Journal of Computer Assisted Learning*, 30(4), 377-395.
- Jenny, H. & Claire, S. (2008). Developing mathematical reasoning through games of strategy played against the computer. *International Journal for Technology in Mathematics Education*, 15(2), 59-72.
- Ju, E., & Wagner, C. (1997). Personal computer adventure games: Their structure, principles, and applicability for training. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, 28(2), 78-92.
- Kennewell, S., & Morgan, A. (2006) Factors influencing learning through play in ICT settings. *Computers & Education*, 46(3), 256-279.
- Kline, P. (1994). *Intelligence: The Psychometric view*. New York, NY: Routledge.
- Kozhevnikov, M. (2007). Cognitive styles in the context of modern psychology: Toward an integrated framework of cognitive style. *Psychological Bulletin*, 133(3), 464-481.
- Krulik, S., & Rudnick, J. A. (1993). *Reasoning and problem solving: A Handbook for elementary school teachers*. Boston, MA: Allyn & Bacon.
- Lee, C. H. M., Cheng, Y. W., Rai, S., & Depickere, A. (2005). What affect student cognitive style in the development of hypermedia learning system? *Computers & Education*, 45(1), 1-19.
- Leighton, J. P., & Sternberg, R. J. (Eds.). (2004). *The Nature of reasoning*. New York, NY: Cambridge University Press.
- Li, M.-C., & Tsai, C.-C. (2013). Game-based learning in science education: A Review of relevant research. *Journal of Science Education and Technology*, 22(6), 877-898.
- Lin, Y. L., Chuang, T. Y., Su, S. H., & Liu, C. C. (2011, May). *The Content analysis of cognitive style in digital game: A Case of Machinarium*. Paper presented at the Proceedings of the 15th Global Chinese Conference on Computers in Education (GCCCE 2011), Hangzhou, China.
- Liu, E. Z. F., & Lin, C. H. (2009). Developing evaluative indicators for educational computer games. *British Journal of Educational Technology*, 40(1), 174-178.
- Lohman, D. F., & Lakin, J. (2011). Reasoning and intelligence. In R. J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge Handbook of Intelligence* (2nd ed.) (pp. 419-441). New York, NY: Cambridge University Press.
- Lomberg, C., Kollmann, T., & Stöckmann, C. (2017). Different styles for different needs—The Effect of cognitive styles on idea generation. *Creativity and Innovation Management*, 26(1), 49-59.
- Lugli, L., Ragni, M., Piccardi, L., & Nori, R. (2017). Hypermedia navigation: Differences between spatial cognitive styles. *Computers in Human Behavior*, 66, 191-200.
- Mampadi, F., Chen, S. Y., Ghinea, G., & Chen, M. P. (2011). Design of adaptive hypermedia learning systems: A Cognitive style approach. *Computers & Education*, 56(4), 1003-1011.
- Mather, N. (1986). Fantasy and adventure software with the LD student. *Journal of Learning Disabilities*, 19(1), 56-58.
- Mefoh, P. C., Nwoke, M. B., Chukwuorji, J. C., & Chijioke, A. O. (2017). Effect of cognitive style and gender on adolescents' problem solving ability. *Thinking Skills and Creativity*, 25, 47-52.

- Messick, S. (1984). The Nature of cognitive styles: Problems and promises in educational research. *Educational Psychologist*, 19, 59-74.
- Moreno, J. (2012). Digital competition game to improve programming skills. *Educational Technology & Society*, 15(3), 288-297.
- Nickerson, R. S. (1991). Modes and models of informal reasoning: A Commentary. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal Reasoning and Education* (pp. 291-309). Hillsdale, NJ: Erlbaum.
- Parkinson, A., & Redmond, J. A. (2002, February). *Do cognitive styles affect learning performance in different computer media?* Paper presented at the ACM SIGCSE Bulletin, New York, NY.
- Piaget, J., & Inhelder, B. (2000). *The Psychology of the child*. New York, NY: Basic Books.
- Raven, J. C. (1936). *Mental tests used in genetic studies: The Performances of related individuals in tests mainly educative and mainly reproductive* (Unpublished master's thesis). University of London, London, United Kingdom.
- Riding, R., & Cheema, I. (1991). Cognitive styles: An Overview and integration. *Educational Psychology*, 11(3), 193-215.
- Riding, R. J., & Sadler-Smith, E. (1997). Cognitive style and learning strategies: Some implications for training design. *International Journal of Training and Development*, 1(3), 199-208.
- Riding, R., & Rayner, S. (2013). *Cognitive styles and learning strategies: Understanding style differences in learning and behavior*. New York, NY: Routledge.
- Rosser, R. A. (1994). *Cognitive development: Psychological and biological perspectives*. Boston, MA: Allyn and Bacon.
- Sadler-Smith, E. (2001). The relationship between learning style and cognitive style. *Personality and Individual Differences*, 30, 609-616.
- Salih, A., & Erdat, C. (2007). The Effects of students' cognitive styles on conceptual understandings and problem-solving skills in introductory mechanics. *Research in Science & Technological Education*, 25(2), 167-178.
- Spitz, L. (1979). Vomiting after pyloromyotomy for infantile hypertrophic pyloric stenosis. *British Medical Journal*, 54(11), 886.
- Sternberg, R. J. (1986). Toward a unified theory of human reasoning. *Intelligence*, 10, 281-314.
- Tamaoka, K. (1985). *Historical development of learning style inventories from dichotomous cognitive concepts of field dependence and field independence to multi-dimensional assessment*. (ERIC Document Reproduction Service No. ED 339729).
- Thomas, P. R., & McKay, J. B. (2010). Cognitive styles and instructional design in university learning. *Learning and Individual Differences*, 20(3), 197-202.
- Volkova, E. V., & Rusalov, V. M. (2016). Cognitive styles and personality. *Personality and Individual Differences*, 99, 266-271.
- Wilhelm, O. (2005). Measuring reasoning ability. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of measuring and understanding intelligence* (pp. 373-392). Thousand Oaks, CA: Sage Press.
- Witkin, H. A., Moore, C. A., Goodenough, D. R., & Cox, P. W. (1977). Field-dependent and field-independent cognitive styles and their educational implications. *Review of Educational Research*, 47(1), 1-64.



# Flipped Classroom in the Educational System: Trend or Effective Pedagogical Model Compared to Other Methodologies?

Héctor Galindo-Dominguez

Facultad de Educación y Psicología, Universidad Francisco de Vitoria, Spain // hector.galindo@ufv.es

(Submitted October 29, 2020; Revised December 20, 2020; Accepted March 16, 2021)

**ABSTRACT:** Flipped Classroom methodology is gaining relative importance as time goes by, in part due to the spreading and accessibility of technological resources in the educational field. Nonetheless, the effectiveness of this methodology is still being discussed. In this sense, the aim of this study is to analyse whether flipped classroom methodology is a more effective methodology than other methodologies. For this purpose, a systematic review was carried out, considering as valid studies those that had a pre-post and a control group. Based on a total of 61 studies ( $n = 5541$  students) from 18 databases, results revealed that Flipped Classroom methodology is more effective than other methodologies in terms of learning achievement, in secondary and higher education, and it could be more beneficial than other methodologies in other constructs as motivation, self-efficacy, cooperativeness and engagement, among others. In primary education, findings revealed that Flipped Classroom could be as effective as other methodologies with regard to learning achievement, and other construct, such as self-concept and social climate. Depending on the educational stage, the effect size of differences was between 1.36 to 1.80 times larger in the case of Flipped Classroom group in comparison with control group. Based on these results, the Flipped Classroom could be more beneficial in comparison with traditional methodologies that are mainly used in higher education. However, it would not more beneficial in other educational stages where traditional approaches are not commonly used, such as in primary education.

**Keywords:** Flipped classroom, Primary education, Secondary education, Higher education, Effectiveness

## 1. Introduction

Flipped Classroom methodology is defined as a methodology in which the more practical part of the class (e.g., activities and problem solving), and traditionally done by students outside class, is moved into the classroom session; while what traditionally was done in class (e.g., presentation of information and information transmission teaching) is moved outside and prior to the class (Låg & Grøm, 2019). Flipped Classroom, correctly applied, could be considered as an active learning methodology as it is an instructional method that engages students in their learning process (Bishop & Verleger, 2013; Prince, 2004). The term Flipped Classroom is relatively new within the educational field (Berrett, 2012). However, it is not a novel teaching methodology since over the last decade analogous terms, such as inverted classroom (Lage, Platt & Treglia, 2000), just-in-time teaching (Novak, 2011) and inverted learning (Davis, 2013) have been studied in the literature to explain this approach, and which emphasize students' work before attending a class (Hung, 2015).

From previous systematic reviews it has seen how the quantity of Flipped Classroom studies was significantly higher in Higher Education than in other educational stages (Uzunboyly & Karagözlü, 2017). Furthermore, the quantity of studies based on Flipped Classroom and performed in Higher Education represent between the 52% and 79% of the total quantity of studies around Flipped Classroom, in comparison with studies carried out in Primary Education that represent around 6% to 7%, and in Secondary Education, it represents from 6% to 8% of the total quantity of studies around the Flipped Classroom (Cheng, Hwang, & Lai, 2020; Uzunboyly & Karagözlü, 2017).

From previous literature, the vast majority of Flipped Classroom interventions are done in the same way (Cheng et al., 2020). Firstly, out-of-class, students access through a learning platform or system, where all the resources are uploaded. This platform has the aim of fostering the learning process around these resources. Secondly, in-class, there are 3 main strategies used: issue discussions, practicing or performing exercises, and group projects (Cheng et al., 2020).

With regard to the matter studied, previous reviews revealed that the main aim of a great number of previous Flipped Classroom papers is to discover the effectiveness of this methodology in terms of academic performance, far from other affective constructs like motivation or satisfaction (Cheng et al., 2020; Galindo-Domínguez & Bezanilla, 2018; Galindo-Domínguez & Bezanilla, 2019).

Results are mostly in favour of Flipped Classroom methodology. In previous reviews, the vast majority of studies revealed positive or, at least, neutral direct or indirect effects, mainly on academic performance and satisfaction with the experience (Chen, Hwang & Lai, 2020; Galindo-Domínguez, 2018; Galindo-Domínguez & Bezanilla, 2019; Låg & Grøm, 2019; O’Flaherty & Phillips, 2015). However, this information has not been contrasted with other methodologies. Nevertheless, other meta-analysis has shown that, in spite of the large number of studies that revealed small positive effects when applying Flipped Classroom methodology, some also pointed out negative effect sizes for the flipped classroom condition (Chen, Ritzhaupt, & Antonenko, 2019).

Chen’s et al. (2018) meta-analysis studied the impact of Flipped Classroom interventions on university students comparing their pre-post academic performance values. Their findings revealed how Flipped Classroom interventions had a statistical significant impact within university students ( $n = 7$ ;  $p < .005$ ), especially in those students enrolled in the health area. This conclusion was also reached by other meta-analysis such as Låg & Grøm’s (2019) study, but, on the other hand, Gillete’s et al. (2018) meta-analysis did not find significant differences between Flipped Classroom and traditional lecture methodology. This is the reason why some authors claim that there is a lack of evidence for the efficiency of Flipped Classroom methodology (e.g., Betihavas, Bridgman, Kornhaber, & Cross, 2016).

However, most of the previous studies which compare pre and post values with a control group, show the effectiveness of this methodology against traditional methodology in terms of different constructs in Higher Education (e.g., Kurt, 2017; Lin & Hwang, 2018; Chang, Kao & Hwang, 2020; Chyr et al., 2017), but some discrepancies appear in Secondary Education (e.g., Kumar, Chang & Chang, 2016; Wei et al., 2020; Gómez-García, Sellés, & Ferriz, 2019) and primary education (e.g., Galindo-Domínguez, 2019a; Galindo-Domínguez, 2019b; Ferriz, Sebastián, & García, 2017; Cheung & Chen, 2020).

Although this review can make an approximation to the impact of this methodology, the results are still incomplete. Specifically, as the literature points out, there is a clear scarcity of previous evidence of meta-analyses and systematic reviews which compares the effectiveness of flipped classroom with other methodologies (Chen et al., 2019).

The justification for this study has its origin in that previous meta-analyses and systematic reviews are focused only on the impact of flipped classroom experiences, but they do not follow a specific selection of the research design of flipped classroom studies (Uzunboyly & Karagözlü, 2017; Cheng et al., 2020). Therefore, it is important, in order to go in depth, to select, analyse and compare studies of at least, pre and post phases, with a control group. This type of design could be the most beneficial one for having a closer and accurate sight of the effectiveness of the flipped classroom, in such a way that it permits to compare the effectiveness of an intervention with the passage of time and according to a specific group.

In addition, there is no previous evidence of meta-analysis or systematic reviews which compare the effectiveness of the flipped classroom depending of the educational stage. In this sense, previous meta-analyses and systematic reviews do not differentiate the educational stage of students, and in fact, this could be a critical factor to take into account when applying a Flipped Classroom intervention (e.g., Uzunboyly & Karagözlü, 2017; Cheng et al., 2020). Due to the fact that the psychosocial characteristics of students are different in each of the different stages, this differentiation may have consequences on the effectiveness of a certain methodology.

Furthermore, there is a significant gap on the constructs studied in previous meta-analyses and systematic reviews, in a way that the vast majority of them are focused on the effectiveness of the Flipped Classroom only considering students’ learning achievement (Chen et al., 2018; Galindo-Domínguez, 2018; Gillete et al., 2018; Låg, & Grøm, 2019). In this sense, this study also analyses the impact of other cognitive, affective and social constructs. It is important to compare the effectiveness of educational methodologies in order to be able to provide teachers with as much information as possible, and thus, base their pedagogical practice on scientific evidence and make justified decisions. This does not necessarily mean that what they do will work, but it means that they already have prior scientific support on which to rely to try to select the best available option, and therefore, allow them to improve their pedagogical practice. It is important to take into account the integral development of the student as it is one of the objectives of the 21st century education, collected in the curricula and educational laws of several countries (for instance, Spain, France, and the United Kingdom). It is for this reason that it is necessary to study the potential of this methodology not only from its cognitive aspect, but also from emotional-affective and social aspects.

## 2. Methodology

### 2.1. Objective

The aim of this paper is to carry out a systematic review considering research that study the impact of Flipped Classroom methodology within the educational system. More specifically, this study analyses the effectiveness of Flipped Classroom interventions in comparison with control methodologies. For this purpose, this research will answer the following questions:

- Is the Flipped Classroom methodology as effective as other methodologies?
- Is the Flipped Classroom methodology as effective at the different educational system stages?
- If not, at what educational stages is the Flipped Classroom methodology most effective?

### 2.2. Documentary search

In order to achieve the objective of this study, certain national and international databases were used. In this case, an exhaustive search was performed in the databases of *Web of Science*, *Scopus*, *InCites*, *ProQuest*, *ScienceDirect*, *SpringerLink*, *Psyc*, *EBSCOHost*, *ACM*, *IEEE Xplore Digital Library*, *Emerald Insight*, *DOAJ*, *Google Scholar*, *PubMed*, *ResearchGate*, *SciELO* and *Dialnet*. Within these databases, the search for documents did not have a starting date but had a deadline of October 2020. These databases were selected because they are the databases that collect the scientific journals with the highest quality and impact at national and international levels. In this specific case, as it is a systematic review, the main interest in the selection of solid and quality studies justifies the usage of these databases.

The search looked into the, the possible crosses between the keywords *Flipped Classroom*, *Flipped Learning* and *Flipped*, with *control group* and *post* were done. All these keywords were also translated and used in the same way in Spanish.

### 2.3. Inclusion criteria

After this first search, a wide range of potential documents was obtained ( $n = 150$ ). Nevertheless, some of them were rejected because they did not fit the inclusion criteria followed for this systematic review. The followed criterion was the next one:

- Accessibility: All results obtained from selected databases were taken into account. Those studies, regardless the format (paper, proceedings...), that were not accessible for the author had to be excluded ( $n = 5$ ).
- Topic: With regard to the topic, only studies focused on Flipped Classroom methodology were taken into account. In this sense, 15 studies were not included in the analysis.
- Sample: It was a required condition that the Flipped Classroom was within the educational system. From this criterion 2, studies focused on the labour field were rejected.
- Construct studied: All cognitive, social and emotional constructs were studied. From this analysis, studies of satisfaction with the experience ( $n = 7$ ) were excluded due to the fact that the focus was to analyse psychological constructs, which were widely studied and consolidated in the scientific literature, as they could provide higher quality and accurate information.
- Methodology: All included studies had to follow a quantitative methodology in order to permit comparisons and extract conclusions based on data. Consequently, 19 studies were rejected from this systematic review as they used qualitative methods or they were meta-analysis.
- Design: In order to permit solid comparisons, it was required to select studies with a control and experimental group, as well as studies with a pre and a post phase. Hence, those studies without a control group ( $n = 16$ ) and/or without a pre and post phase ( $n = 22$ ) were excluded from the analysis. Finally, there were some studies that mixed the methodology of the experimental and the control group, that is, what was at first the control group swapped to the experimental group, and vice versa. These studies ( $n = 2$ ) were not included as they would significantly complicate drawing conclusions.
- Language: Studies that were not in Spanish or English were excluded.
- Once the studies passed through the explained criteria, a total of 61 research studies were selected, 58 in English and 3 in Spanish. The process of this analysis was performed by an adapted PRISMA flow diagram (Moher et al., 2009), as gathered in Figure 1. These studies analysed the impact of Flipped Classroom based on different constructs. 31 of them analysed the data by means of the repeated measures ANOVA, and 30 of them analysed the data by means of an analysis of covariance after observing that in the pre phase the

control and experimental groups did not show significant differences. The results of this study are based on these 61 research studies.

Noteworthy to mention that in those multidimensional constructs that did not provide an overall score, the arithmetic mean amongst the different dimensions were done.

Finally, as the countries in which the studies were carried out have different educational systems, and therefore, different ages for each educational group, the grouping mode for this analysis could be affected. That is why the studies were grouped with respect to educational stages as follows: (1) Primary Education: it was considered as primary education those students from 6 to 12 years old; (2) Secondary Education: It was considered as secondary education those students from 12 to 16; (3) It was considered as university education those students beyond 18 years old. No studies were found with the selected criteria for students aged 16 to 18 years. This post-high school stage is called differently depending on the country. To name a few, in Spain it is known as Baccalaureate, in the United Kingdom the A level of the General Certificate of Education (GCE), in Saudi Arabia Tawjahiya or in Belgium Higher Secondary Education.

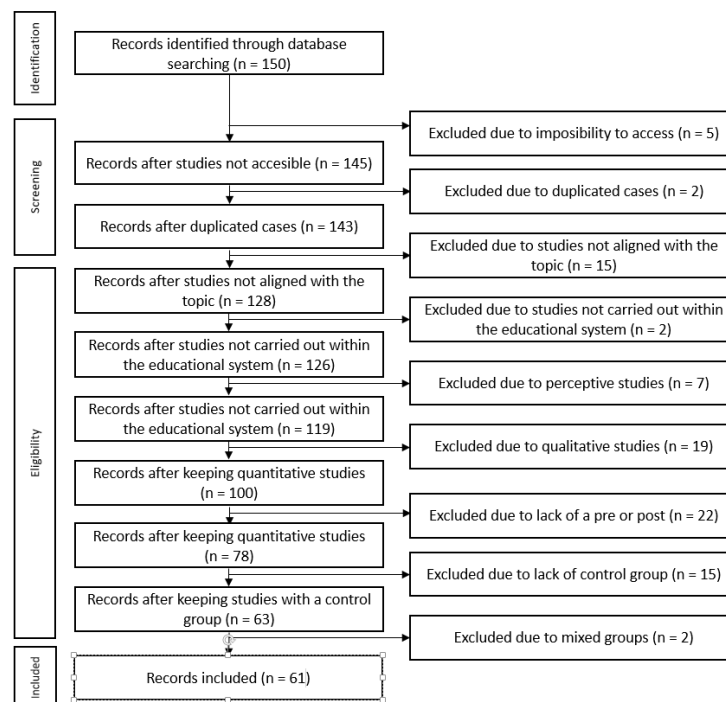


Figure 1. PRISMA Flow diagram of the inclusion criteria

## 2.4. Data analysis

After gathering the research studies in line with the initial criteria, their content was analysed. For this purpose, SPSS Statistics 23.0 was used. Through this software some interesting variables from all the selected studies were gathered (educational stage, intervention's duration, construct studied, sample of the control and experimental group, type of methodology followed in the control and experimental group, and means and standard deviations of the control and experimental group for the pre and post phase). To encode the constructs, the objective of each study was analysed, and in those studies in which more than one construct was analysed, more than 1 row was used in the database (1 row for each construct studied in the investigation). Finally, an inductive analysis according to the typology of the constructs studied was carried out, allowing to classify the total of constructs into: (1) cognitive constructs, those related to cognitive intelligence. Specifically, academic performance was found as the main cognitive construct; (2) affective-emotional constructs, that is, those constructs that are mainly related to the management and understanding of the individual's internal emotions, and which have an impact on their well-being and productivity. Specifically, self-concept, motivation, engagement, self-direction, metacognition, self-regulation and anxiety; (3) social constructs, namely, those internal or external constructs that have a high impact on the processes of interaction with other individuals. Specifically, competitiveness, cooperativeness and social climate.

An attempt was made to achieve the highest reliability in this data analysis process by establishing in advance a series of variables, which can be seen in Tables 1 and 2. In the case of Table 1, where the differences between the pre and post data for the control group and the experimental group are collected, the selected studies did not use an instrument with validated and reliable psychometric properties. In the case of Table 2, all the studies make use of instruments previously validated and consolidated by previous theories. It should be highlighted that in all cases the study was carried out with convenience samples.

After the database was consolidated, a descriptive analysis was carried out through the means and standard deviations of all selected studies. Then, Cohen's  $d$  was calculated by its formula  $d = (M_2 - M_1) / SD_{\text{pooled}}$ .

Finally, in order to perform a comparison between the control and the experimental group, the repeated measures ANOVA was carried out. Firstly, the repeated measures ANOVA was carried out considering pre and post phases of control group as within-subject factors and educational stage as between-subject factor. Then, the same procedure was carried out with the experimental group. In these analyses, the differences between pre and post, as well as the impact of the educational level along time were studied. This analysis provided 2 different plots, one for each group. That is the reason why an external graphic software was used to combine both plots in one in order to facilitate the interpretation between groups.

### 3. Results

Firstly, the impact of Flipped Classroom methodology on academic performance was studied. As gathered in Table 1, a total of 31 studies were included in this analysis: 3 studies focused on primary education, 9 studies focused on secondary education and 19 studies focused on university education. Other studies also analysed the impact of Flipped Classroom on academic performance, but used an analysis of Covariance, thus, making it impossible to introduce and compare the data with the information shown in Table 1. Nonetheless, these studies are used for justification or rejection of the findings.

It is important to highlight that all the control groups were grouped under the name of "control methodologies," which according to the authors of these studies, mainly used a traditional methodology. Nevertheless, in practice, it is likely that to a lesser degree, other types of methodologies not indicated in the "description of the intervention" section of the different studies were used.

The overall results from this analysis and confirmed through the repeated measures ANOVA showed that, regardless the educative stage, the post values were higher than the pre values for both: control group ( $p = .013$ ) and experimental group ( $p = .003$ ). The interaction between time and educational level resulted in non-significant differences in both groups, experimental group ( $p = .680$ ) and control group ( $p = .456$ ), stating that, regardless the educational stage of the sample, all of them improve their academic performance. Nonetheless, these results required further analysis, in order to detect possible significant differences amongst the different educational stages.

At this point, it is required to analyse and compare the impact of the group (experimental and control) at each educational stage.

As illustrated in Figure 2, it is shown that in Primary Education ( $n = 3$ ), Flipped Classroom methodology is not more significant in the control group's methodology as experimental ( $\bar{x}_{\text{pre}} = 6.51 \pm 2.92$ ;  $\bar{x}_{\text{post}} = 14.34 \pm 5.49$ ;  $p = .147$ ) and the control group ( $\bar{x}_{\text{pre}} = 6.24 \pm 3.32$ ;  $\bar{x}_{\text{post}} = 12.80 \pm 5.55$ ;  $p = .132$ ) obtained non-significant differences from pre phase to post phase. These results show that, regardless the methodology used, the impact on the academic performance is low. However, these results should be taken carefully as only 2 studies could be analysed. Some causes with regard this piece of information are discussed later on. This idea is supported by other primary education-focused studies which carried out a different methodology. More specifically, in Ferriz's et al. (2017) study it was shown how there was not a statistical significant difference between students who learnt through Flipped Classroom methodology and students who studied through the conventional methodology on their academic performance.

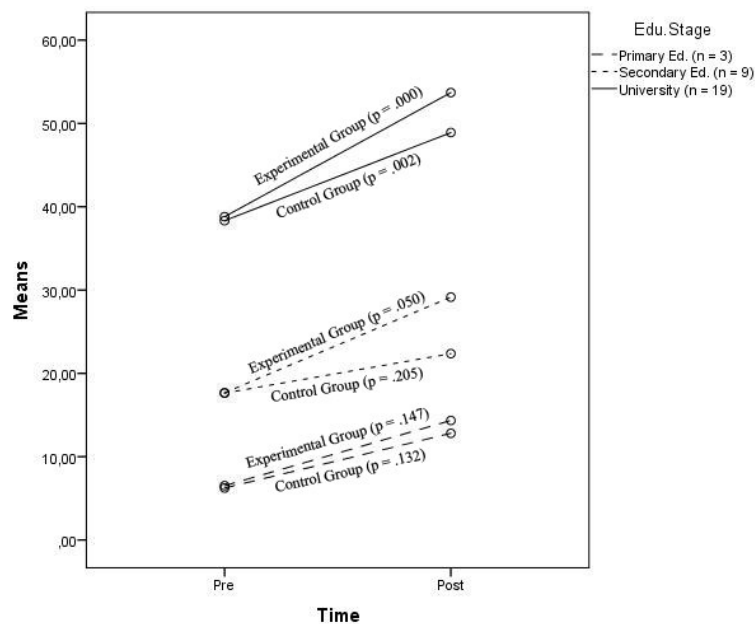


Figure 2. Impact of control and experimental groups' interventions on academic performance

With regard to secondary education ( $n = 9$ ), the results reveal how the control group did not improve their academic performance significantly with their intervention ( $\bar{x}_{pre} = 17.62 \pm 18.42$ ;  $\bar{x}_{Post} = 22.36 \pm 23.65$ ;  $p = .205$ ) in comparison with the students from the experimental group, who improved significantly their academic performance with their Flipped Classroom intervention ( $\bar{x}_{pre} = 17.67 \pm 18.68$ ;  $\bar{x}_{Post} = 29.16 \pm 30.44$ ;  $p = .050$ ). This information matches with the conclusion of Djamaa's (2020) study, which revealed a significant benefit for secondary school students who learnt with Flipped Classroom approach in comparison with students who learnt with a more traditional approach.

Finally, concerning university education, the results reveal how both groups, control ( $\bar{x}_{pre} = 38.32 \pm 24.62$ ;  $\bar{x}_{Post} = 48.90 \pm 27.98$ ;  $p = .002$ ) and experimental ( $\bar{x}_{pre} = 38.80 \pm 24.96$ ;  $\bar{x}_{Post} = 53.70 \pm 30.32$ ;  $p = .000$ ), improved significantly their academic performance with their intervention. These results are also supported by other quantitative studies that could not be included in Table 1, as they used a different type of analysis (an ANCOVA, for instance). In this sense, Mattis (2015), Jian (2019), Wasserman et al. (2017), Chang et al. (2019), and Lai, Ting, and Yuch (2020) supported the idea that Flipped Classroom methodology could be even more beneficial than a more traditional approach to improve university students' academic performance. This does not mean that the control methodologies are not effective (which in fact, as can be seen in Figure 2, the control group students also show improvement), but that the Flipped Classroom methodology could be even better in higher education.

From these results, the Flipped Classroom could be potentially beneficial, especially from secondary education till university education, and equally beneficial than other methodologies used in primary education. If effect sizes between control and experimental group are compared (dividing the effect size of the experimental group by the control group), it can be seen how primary education students ( $n = 3$ ) from experimental group obtained a 1.60 times larger effect size in comparison with students from the control group. Nonetheless, when removing excessive large effect size studies (Elia & Hamaidi; 2018) the experimental group showed 1.36 times larger effect sizes. In the case of secondary school students, the effect size was larger for experimental group students, in a way that the experimental group obtained a 3.67 times larger effect size in comparison with students from the control group. This information should be taken into account with care as there is a study from Mahmoud (2020) with a very large Cohen's  $d$  values. Withdrawing this study, the effect size is still larger (1.80 times) for the experimental group in comparison with the control group. Finally, in the case of university education, the effect size was 1.65 times larger (1.60 times if dismissing Sezer & Abay's study due to a very large Cohen's  $d$ ) for the experimental group than for the control group.

Table 1. Main results of the different studies that had a pre-post and control and experimental group about academic performance ordered by effect size

Authors	Intervention	Control Group (CG)				Experimental Group (EG)				Results
		<i>n</i>	Pre	Post	<i>d</i>	<i>n</i>	Pre	Post	<i>d</i>	
Primary Education										
Elian & Hamaidi (2018)*	3 weeks	22	4.72 ± 2.18	16.27 ± 2.47	4.95	22	4.63 ± 2.12	19.09 ± 1.01	8.70	EG > CG
Jiménez & Domínguez (2018)	N / D	21	3.95 ± 1.50	6.40 ± 1.68	1.53	19	5.03 ± 1.28	8.33 ± 1.08	2.78	EG = CG
Cheung & Chen (2020)	4 weeks	95	10.05 ± 4.09	15.75 ± 2.82	1.62	94	9.89 ± 3.64	15.62 ± 3.84	1.53	EG = CG
Overall		46	6.24 ± 3.32	12.80 ± 5.55	2.7	45	6.51 ± 2.92	14.34 ± 5.49	4.33	
Overall [without outliers]*		58	7.00 ± 2.79	11.07 ± 2.25	1.57	56,5	7.46 ± 2.46	11.97 ± 2.46	2.15	
Secondary Education										
Mahmoud (2020)*	4 months	75	33.33 ± 1.72	34.17 ± 2.47	0.39	73	33.23 ± 1.80	69.98 ± 4.37	10.99	EG > CG
Schmeisser et al. (2018)	40 weeks	22	11.86 ± 5.42	39.45 ± 1.54	6.92	21	9.52 ± 6.63	36.76 ± 7.05	3.98	EG = CG
Eyitayo (2017)	3 weeks	33	5.73 ± 2.75	7.14 ± 2.54	0.53	33	5.12 ± 2.53	10.82 ± 2.44	2.29	EG > CG
Gómez-García et al. (2019)	3 weeks	26	4.52 ± 1.61	6.17 ± 1.55	1.04	30	5.43 ± 1.66	7.85 ± 1.37	1.59	EG = CG
Wei et al. (2020)	5 weeks	44	60.00 ± 23.82	75.52 ± 22.68	0.66	44	60.82 ± 22.93	88.95 ± 20.10	1.30	EG > CG
Namazandost & Shafiee (2018)	N / D	25	12.40 ± 5.12	14.71 ± 5.70	0.42	25	13.38 ± 3.78	18.88 ± 6.72	1.00	EG > CG
Hamdani (2019)	3 months	39	2.26 ± 0.93	2.62 ± 1.03	0.36	38	2.65 ± 0.83	2.99 ± 0.78	0.42	EG > CG
Kumar et al. (2016)	6 weeks	41	9.12 ± 1.26	8.81 ± 2.00	-0.18	41	9.49 ± 1.90	9.97 ± 2.55	0.21	EG > CG
Mustapha (2020)	13 lessons	20	19.08 ± 0.24	12.72 ± 1.98	-4.50	20	19.40 ± 0.24	16.20 ± 3.58	-1.26	EG > CG
Overall		36,1	17.62 ± 18.42	22.36 ± 23.65	0.62	36,1	17.67 ± 18.68	29.16 ± 30.44	2.28	
Overall [without outliers]*		31,2	15.66 ± 18.66	20.89 ± 24.84	0.66	31,5	15.72 ± 18.97	24.05 ± 28.13	1.19	
Higher Education										
Sezer & Abay (2018)*	8 weeks	19	33.52 ± 3.62	61.00 ± 5.28	6.07	19	33.15 ± 4.01	82.10 ± 4.71	11.19	EG > CG
Robert et al. (2017)	25 hours	137	46.90 ± 9.80	86.10 ± 5.00	5.03	137	48.30 ± 10.40	86.00 ± 5.30	4.56	EG = CG
Penichet et al. (2017)	N / D	35	4.26 ± 1.60	6.31 ± 1.57	1.29	29	2.86 ± 1.46	8.31 ± 1.28	3.96	EG = CG
Zheng et al. (2018)	6 months	61	38.13 ± 13.01	76.32 ± 10.56	3.22	76	37.36 ± 13.23	82.30 ± 9.39	3.91	EG = CG
Wyk (2018)	6 months	162	58.30 ± 3.78	69.01 ± 6.71	1.96	209	58.77 ± 3.15	72.15 ± 4.21	3.59	EG = CG
Talan & Gulsecen (2019)	7 weeks	40	15.45 ± 4.95	30.45 ± 5.94	2.74	40	16.40 ± 5.27	33.95 ± 4.65	3.53	EG = CG
Haghighi et al. (2018)	7 lessons	30	27.30 ± 4.87	36.37 ± 5.22	1.79	30	27.80 ± 3.97	42.70 ± 4.85	3.36	EG > CG
Karabatak & Polat (2020)	8 weeks	31	30.42 ± 13.15	45.48 ± 9.55	1.31	35	30.80 ± 6.41	53.11 ± 7.87	3.10	EG > CG
Lin & Hwang (2018)	18 weeks	16	17.12 ± 1.89	19.62 ± 1.20	1.57	33	18.60 ± 1.69	22.12 ± 1.26	2.36	EG > CG
Alsancak & Özdemir (2018)	3 weeks	34	57.20 ± 11.40	72.04 ± 9.63	1.40	32	61.80 ± 10.40	79.41 ± 7.35	1.95	EG > CG
Kazanidis et al. (2018)	12	62	43.13	55.77	1.44	66	43.06	61.46	1.92	EG >

	weeks		± 5.66	± 11.02			± 5.57	± 12.33		CG
Sommer & Ritzhaupt (2018)	15 weeks	31	3.10 ± 1.51	6.55 ± 1.06	2.64	41	3.53 ± 1.87	6.51 ± 1.23	1.88	EG = CG
Zainuddin & Jacqueline (2017)	12 weeks	30	74.43 ± 8.29	79.77 ± 2.67	0.86	31	78.26 ± 9.45	89.64 ± 4.61	1.53	EG > CG
Chu et al. (2019)	5 hours	75	65.33 ± 18.55	75.07 ± 14.55	0.54	76	59.21 ± 18.85	80.92 ± 14.62	1.28	EG > CG
Fan et al. (2020)	6 months	198	5.38 ± 0.66	5.48 ± 0.63	0.15	287	5.39 ± 0.64	5.71 ± 0.56	0.53	EG > CG
Hava (2018)	5 weeks	33	81.45 ± 5.03	82.20 ± 4.84	0.15	26	80.15 ± 8.60	83.30 ± 5.03	0.44	EG = CG
Knežević et al. (2020)	8 weeks	30	11.26 ± 4.78	14.86 ± 2.51	0.94	30	11.03 ± 4.63	12.16 ± 4.35	0.25	EG > CG
Foldnes (2016)	6 months	142	60.70 ± 2.70	50.10 ± 3.90	-3.16	93	62.50 ± 2.70	63.20 ± 4.00	0.20	EG > CG
Cabi (2018)	4 weeks	31	54.84 ± 18.56	56.64 ± 14.79	-0.10	28	58.33 ± 18.98	55.29 ± 16.11	-0.17	EG = CG
Overall		63	38.32 ± 24.62	48.90 ± 27.98	1.57	69,3	38.80 ± 24.96	53.70 ± 30.32	2.59	
Overall [without outliers]*		65,4	38.59 ± 25.31	48.23 ± 28.63	1.32	72,1	39.11 ± 25.64	52.12 ± 30.39	2.12	

Apart from academic achievement, there is a large list of constructs that have also been considered in the literature and are analysed below.

*Table 2.* Main results of the different studies that had a pre-post and control and experimental group about different psychological constructs ordered by effect size

Construct	Authors	Inte.	Control Group (CG)				Experimental Group (EG)				Results
			<i>n</i>	Pre	Post	<i>d</i>	<i>n</i>	Pre	Post	<i>d</i>	
Primary Education											
Self-concept	Galindo-Domínguez (2019)	7 weeks	437	3.95 ± 0.64	3.99 ± 0.64	0.06	385	4.03 ± 0.63	4.06 ± 0.64	0.04	EG = CG
Social climate	Galindo-Domínguez (2019)	7 weeks	437	4.06 ± 0.57	4.04 ± 0.54	-0.03	385	4.10 ± 0.55	4.07 ± 0.50	-0.05	EG = CG
Secondary Education											
Motivation	Ruiz (2016)	4 months	23	45.64 ± 17.00	46.40 ± 17.02	0.04	25	43.00 ± 13.61	66.00 ± 17.95	1.44	EG > CG
Engagement	Ayçiçek, & Yanpar, (2018)	4 weeks	20	11.80 ± 5.56	13.74 ± 5.58	0.34	20	13.84 ± 5.90	16.72 ± 5.76	0.49	EG > CG
Higher Education											
Self-direction	Chyr et al. (2017)	6 months	35	3.16 ± 0.15	3.18 ± 0.22	0.10	34	3.14 ± 0.18	3.30 ± 0.25	0.73	EG > CG
	Hava (2018)	5 weeks	33	105.75 ± 10.69	108.60 ± 12.44	0.24	26	106.73 ± 11.29	110.34 ± 10.03	0.33	EG = CG
Self-efficacy	Kurt (2017)	14 weeks	30	136.07 ± 20.40	155.87 ± 19.13	1.00	32	125.22 ± 27.30	162.72 ± 21.03	1.53	EG > CG
	Chu et al. (2019)	5 hours	75	63.61 ± 16.39	82.15 ± 17.52	1.09	76	62.76 ± 21.66	89.03 ± 15.19	1.40	EG > CG
	Chyr et al. (2017).	6 months	35	3.95 ± 0.40	3.71 ± 0.67	-0.43	34	3.83 ± 0.50	4.39 ± 0.56	1.05	EG > CG
	Namaziandest & Çakmak (2020)	14 weeks	27	23.88 ± 3.86	23.40 ± 3.65	-0.12	31	24.77 ± 3.97	26.09 ± 3.52	0.35	EG > CG
Motivation	Karabatak &	8 weeks	31	3.63 ±	3.53 ±	0.16	35	3.34 ±	3.70 ±	0.62	EG >



	Polat (2020)			0.58	0.64			0.57	0.58		CG
Metacognition	Fan et al. (2020)	6 months	198	3.30 ± 0.49	3.52 ± 0.52	0.43	287	3.41 ± 0.51	3.59 ± 0.52	0.34	EG = CG
Competitiveness	Eon & Rok (2018)	6 months	76	3.73 ± 1.00	3.58 ± 0.99	-0.15	81	3.62 ± 0.89	3.07 ± 0.95	-	EG < CG
Cooperativeness	Eon & Rok (2018)	6 months	76	3.54 ± 0.79	3.63 ± 0.90	0.10	81	3.50 ± 0.82	4.04 ± 0.99	0.59	EG > CG
Self-regulation	Hava (2018)	5 weeks	33	62.36 ± 6.81	62.93 ± 6.23	0.08	26	63.61 ± 6.15	61.50 ± 8.35	-	EG = CG
Engagement	Chyr et al. (2017)	6 months	35	4.16 ± 0.44	4.11 ± 0.49	0.10	34	4.06 ± 0.32	4.31 ± 0.49	0.60	EG > CG
Anxiety	Chang & Koong (2019)	16 weeks	40	3.61 ± 0.84	3.01 ± 0.78	-0.74	45	3.71 ± 0.75	2.89 ± 0.67	-	EG < CG

Firstly, in relation to primary education, besides the information gathered in Table 2, Ferriz et al. (2017) revealed that both, students who applied Flipped Classroom methodology and students who applied a more traditional methodology significantly reduced their discouragement. Hence, based on these studies, the effectiveness of Flipped Classroom in comparison with other methodologies in primary education does not reveal striking findings. Nonetheless, this information should be taken carefully as only 4 pre-post with control group studies using an ANCOVA have been analysed.

Secondly, in relation to secondary education, besides the information gathered in Table 2, Gómez-García et al. (2019) affirm that the Flipped Classroom approach was not a more effective approach than a more traditional methodology in order to improve students' motivation. In this sense, further research about the impact of flipped classroom methodology on social and emotional constructs is required as only 3 studies pre-post with control group using an ANCOVA have been analysed in secondary education.

Thirdly, in relation to higher education, besides the information gathered in Table 2, Jian (2019) and Chang et al. (2019) demonstrated how Flipped Classroom methodology at university fosters students' learning motivation more significantly than traditional approaches. In addition, Beth et al. (2016) and Jdaitawi (2019) make evident that Flipped Classroom methodology could be more beneficial than traditional approaches in order to improve students' self-regulation.

Finally, there are a group of studies not included in table 1 or in table 2, which are not focused on comparing the effectiveness of Flipped Classroom methodology in contrast of other methodologies, but they compare the effectiveness of an adaptation of Flipped Classroom methodology against the conventional Flipped Classroom methodology.

Thus, there is some evidence that the Flipped Classroom methodology complemented with gamification (Aşıksoy, 2018), Reflective thinking-promoting mechanisms (Chen, 2019), RSI (Recognize, Summarize, Inquire) approach (Chang et al., 2020), KM (Knowledge Management) models (Thongkoo, Panjaburee, & Daungcharone, 2019) and Collective issue-quests systems (Chen & Hwang, 2019) could provide a significant improvement on university students' academic performance in comparison with the conventional Flipped Classroom methodology. The same happens in the case of motivation (Liu, Sands-Meyer, & Audran, 2019; Aşıksoy, 2018), self-regulation (Chen & Hwang, 2019), self-efficacy (Liu, Sands-Meyer, & Audran, 2019), and self-concept, critical thinking and problem-solving skills (Chang et al., 2020) improving more significantly these constructs on students applying the adaptation of Flipped Classroom in comparison with students applying a more conventional approach of Flipped Classroom.

## 4. Discussion

The main objective of this study has been to explore the effectiveness of Flipped Classroom methodologies in comparison with other teaching methodologies along the different stages of educational system.

As observed from the repeated measures and Figure 2, findings reveal that the Flipped Classroom could be more beneficial than control methodologies when applied to Secondary and Higher Education students, and equally beneficial than control methodologies when applying it to Primary Education students. These results are partially coherent and complementary with previous meta-analyses and systematic reviews (Chen et al, 2018; Galindo-Domínguez, 2018; Låg & Grøm, 2019) and contrary to other meta-analyses (Gillete et al., 2018). Based on this conclusion, some considerations should be taken into account.

Indeed, the main benefit often commented around the flipped classroom is that students who use this methodology are more prone to develop higher order skills under teacher guidance and peer support, due to the fact that in-time class is more focused on cooperative learning and practical tasks (Berrett, 2012). This change could permit teachers to develop in-class high order thinking skills, based on Bloom's (1984) taxonomy, and to establish a prior autonomous, but guided preparation before class working on low order thinking skills of Bloom's taxonomy (Hung, 2015). However, it is important to highlight other potential benefits over traditional teaching models cited in the literature, like a more personalized teaching and learning process (O'Flaherty & Phillips, 2015), a better management and organization of class time (Herreid, Schiller, Herreid, & Wright, 2014), and an improvement of the responsibility of students for their own learning process (O'Flaherty & Phillips, 2015). Nonetheless, critical reviews of the Flipped Classroom have revealed that there could be problems and future challenges related to this methodology. In this way, Lo and Hew (2017) highlight as negative points (1) that students could not be satisfied after using this methodology because they are not familiar with it and are not used to the routine or procedure it involves; (2) that students believe that the videos are very long and / or cannot pay enough attention when watching them. This may be due to the boredom and passivity that they can generate; (3) that certain students require clearer instructions from the teacher to work the practical part of the lesson in class; (4) that, like homework, activities before class take time and this makes students be overwhelmed by work at home; (5) that students cannot ask their doubts immediately during and after viewing the videos.

Firstly, it is thought that the effectiveness of the Flipped Classroom methodology is linked to the autonomy and responsibility of the student, as students are required to be autonomous for the preparation of the class. From teachers' view, it should be highlighted that, due to students' level of maturity and the disparity of the level of maturity amongst students, there are difficulties when giving primary school students a great deal of freedom of choice, and therefore, hand over to them the management of their learning process (Admiraal, Nieuwenhuis, Kooij, Dijkstra, & Cloosterman, 2019). This could be one of the main reasons why, despite the fact that primary education teachers foster students' autonomy, it is complex to develop a totally autonomous learning processes in children. In addition, in some cases, such as self-concept among others, it should be taken into account that these constructs are considered stable constructs and require large periods of time to modify them (Galindo-Domínguez, 2019b). It may be that for this reason, no significant differences have been found. Focusing on Higher Education, it is true that the literature has emphasized that having the responsibility of one's own learning is, in some cases, more demanding and more frustrating when there is an obvious lack of structure and direction (Boud, 1995; McKay & Emmison, 1995). However, one of the main aims university teachers attempt to foster in their students, differently from school pupils, is to actively pursue their own autonomy in their learning progress (Scoot, Furnell, Murphy, & Goulder, 2015; Thomas, Hockings, Ottaway, & Jones, 2015). In addition, data has shown how final-year university students tend to have higher levels of their own progression and learning than previous year students (Brown, 2007). In this sense, autonomy in students' learning process could be a clear factor that could have a significant impact on carrying out Flipped Classroom interventions, as students are required, among other activities, to read documents, watch videos, connect to the internet, pay attention to their tasks.

Secondly, it is thought that the effectiveness of Flipped Classroom methodology is linked to the accessibility to digital resources and the presence of a medium-high digital competence. Research from last decade (Frederick, 2002) has shown how OECD countries were divided into two groups, based on the accessibility of children to ICT. The first group included highly developed OECD members. This group presents high ICT access rates for children, providing them with an Internet connection and digital resources in schools that facilitate their access to the net. Nevertheless, there was still a divided line in terms of accessibility to Internet at home caused, mainly by socioeconomic factors, such as parents' income. Previous literature has shown that this is the main dividing factor (UNICEF, 2017). The second group included the least developed countries. These groups had not yet provided ICT access to their children at school or through other means. The present context, however, has improved in such a way that from 2006 to 2015, the percentage of children from OECD countries who had access to the Internet at home had been significantly increased up to a 95%. Nonetheless, this situation is not equal for countries like Mexico and Peru, where only one out of two students have access to the Internet at home (OECD, 2017), and this figure is even worse in low-income countries, like Bangladesh and Zimbabwe, where only 1 out of 20 children under 15 year old has access to the internet (UNICEF, 2017). Based on these findings, this fact means a limitation to the Flipped Classroom's methodology as the resources require the need the internet for out-of-class preparation. The lack of access to these technological resources could be more notorious among younger students than among older students who, due to their autonomy and possibility to having a wage that permit them to buy these devices, while the former could have more difficulties in accessing the Internet and having a quality equipment.

Thirdly, it is thought that in Primary Education a wide variety of methodologies are commonly used, like projects or problem-solving in comparison with other educational stages, where the traditional lecture is still one

of the most common methodologies used. Furthermore, previous studies have revealed that at university, around 70% of the activities in which teachers are engaged consist of traditional lectures to students and the most common methodologies used are not active methodologies (Rutkienė, & Tandzegolskienė, 2015; Schmidt, 2010). Therefore, introducing an innovative methodology correctly, like Flipped Classroom, could lead to positive results. This has been previously discussed in the literature, pointing out the potential benefits of active methodologies, such as cooperative learning, experimental learning, and innovative usage of new technologies in education, against a more traditional and passive learning style (e.g., Khan, 2008; Pedró, 2007; Uddin & Khan, 2018). On the contrary, in primary education, despite the fact that Flipped Classroom studies have considered the control group as the group that based their intervention on a traditional methodology, against all odds, it is extremely difficult to find a traditional class in these stages. As a large number of studies have pointed out (Buljubašić & Petrović, 2014; Skutil, Havlíčková, & Matějčková, 2015), in primary education and even, in secondary education, there is a wide variety of methodologies that are commonly used, such as cooperative learning, experiential learning, problem solving, presentations, mind maps, games and simulations, to name a few. In this sense, it is possible that Flipped Classroom methodology could not be as effective as other methodologies within primary education students, when other active learning methodologies are used. In addition, it should be taken into account what Låg and Grøm (2019) claim regarding the novelty of this methodology. In fact, as it is a recent teaching method, first-time usage of new methods may be more prone to unexpected obstacles due to teachers' and students' inexperience. Hence, it would be reasonable to expect possible significant improvements in comparison with other methodologies in the future.

Even so, these results have several theoretical and practical implications that should be highlighted. Concerning theoretical implications, these results reveal the theories and basis behind the Flipped Classroom methodology, as being, at least, equally effective as other teaching methodologies. This means that future studies could gradually improve this methodology, for example, unifying a series of indicators or models that would function as a reference to apply effective interventions in the Flipped Classroom. In addition, having shed some light on the effectiveness of this methodology, it could help teachers to justify their teaching processes based on more scientific evidence. Thus, it may be possible to create impact teaching programs based on this methodology and continue assessing its effectiveness with the passage of time.

Another important idea is that, the flipped classroom is generally compared with other control methodologies (mainly traditional methodologies). Nonetheless, it has been gradually seen how flipped classroom adaptations are being compared with a more traditional flipped classroom model. This could be an interesting future research line, as the results would allow the scientific community to know which complementary methodologies work in a better way than the conventional flipped classroom.

Finally, in spite of the fact that this systematic review has been performed to the best of our possibilities, this study has some limitations that should be taken into account when interpreting these results.

Firstly, due to the complexity of pedagogical practices, the interventions from the experimental group and the control group could have varied. In this sense, despite the fact that the data has been clustered under the Flipped Classroom methodology tag, in practice, teachers could have interpreted this methodology in a different way, and they could have implemented it in a different way. The same phenomenon happens in the case of the control group, which, in some cases, it is classified under the Control methodology tag, when in practice teachers could have developed different practical activities, which could bias the conclusions of this study. In this sense, it is important for future studies to try to provide a more in-depth description of the interventions carried out in the classroom, both for interventions based on the Flipped Classroom (duration of the intervention, methodologies used, sequencing followed, subjects in which it has been intervened, and so on), and those based on other methodologies, and thus, be able to make the strongest possible groupings. In the case of this research, there have been cases in which it has been impossible to know their duration, or that the duration provided has been so short that it would be difficult to show solid changes. In this sense, it would be interesting for future research to propose interventions of a longer duration in time (of some months or even of some years) that would allow to attribute a greater causal relationship of the results to the methodology used.

Secondly, there are some important variables that have not been considered as they are not described within the different studies. In this sense, personal variables like teacher's expertise in the Flipped Classroom or contextual variables, such as the impact of the socioeconomic context, could have a significant impact on the results. In this sense, future studies should try to provide more contextual information about the intervention, and which may allow the researcher to clearly analyse each study with as many significant variables as possible.

Thirdly, it should be highlighted how this study includes studies until October 2020. Recently, there is an important interest that this topic, and this is reflected in the scientific and educational field, having as a result an

exponential amount of research around the subject. In this sense, it is possible that from the new studies could have appeared before the publication of this article.

Fourthly, it should be taken into account that the conclusions of this study are based on a small number of studies available in the current literature. That is why, future studies could repeat the same study with the same methodology in order to compare and contrast the findings of those studies with the results of the present research.

Fifthly, it is necessary to take into account the sample selection procedure followed in the studies. It should be remembered that the sample selection method used is always a non-probabilistic method, and which on certain occasions can lead to certain limitations, such as the lack of representation of certain groups over the total population. It is a complex limitation to overcome, but it would be interesting if future studies could try to carry out research following completely random sampling system. However, despite these limitations, this study has some strengths. For instance, it has been the first systematic review that compares the effectiveness of Flipped Classroom methodology in comparison to other methodologies regarding numerous constructs beyond academic performance. In addition, the results have allowed to know how the effectiveness of this methodology could vary depending on the educational stage taken into account, and this could be a significant contribution to the scientific community.

Lastly, it is noteworthy to comment amongst the limitations how the content analysis process was performed solely by the author of the article. In order to avoid subjective biases, future studies could attempt to carry out this content analysis process with the presence of more than 1 researcher.

It is clearly important to continue investigating the effectiveness of active methodologies insofar teachers want to base their practice on scientific evidence, leaving aside educational fashions and trends without scientific basis, and thus, get to know which methodologies are those that work best in a specific context. Moreover, it is also necessary to continue providing and conducting in-depth comparative research to provide teachers with effective tools.

## Acknowledgement

The author would like to express his gratitude to Dr. Donna Fernández for her assistance.

## References

- Admiraal, W., Nieuwenhuis, G., Kooij, Y., Dijkstra, T., & Cloosterman, I. (2019). Perceived autonomy support in primary education in the Netherlands: Differences between teachers and their students. *World Journal of Education*, 9(4), 1-12.
- Aşıksoy, G. (2018). The Effects of the gamified flipped classroom environment (GFCE) on students' motivation, learning achievements and perception in a physics course. *Quality & Quantity*, 52, 129-145.
- Berrett, D. (2012). How "flipping" the classroom can improve the traditional lecture. *The Chronicle of Higher Education*, 58(25), 16-18.
- Beth, A., Hedges, A., Benedict, N. J. & Donihi, A. (2016). Combination of a flipped classroom format and a virtual patient case to enhance active learning in a required therapeutics course. *American Journal of Pharmaceutical Education*, 80(10), 1-8.
- Betihavas, V., Bridgman, H., Kornhaber, R., & Cross, M. (2016). The Evidence for "flipping out": A Systematic review of the flipped classroom in nursing education. *Nurse Education Today*, 38, 15-21.
- Bishop, J. L. & Verleger, M. A. (2013). The Flipped classroom: A Survey of the research. In W. W. Buchanan (Pres.), *120th ASEE Annual Conference and Exposition* (pp. 1-18). ASEE.
- Bloom, B.S. (1984). *Taxonomy of educational objectives*. Boston, MA: Allyn and Bacon.
- Boud, D.J. (1995). *Enhancing learning through self assessment*. London, UK: Kogan Page.
- Brown, J. (2007). Feedback: the student perspective. *Research in Post-Compulsory Education*, 12(1), 33-51.
- Buljubašić, V. & Petrović, A. (2014). Teaching and lesson design from primary and secondary teachers' perspective. *Život i škola*, 60(31), 76-90.

- Chang, B., Chang, C., Hwang, G., & Kuo, F. (2019). A Situation-based flipped classroom to improving nursing staff performance in advanced cardiac life support training course. *Interactive Learning Environments*, 27(8), 1062-1074.
- Chang, C., Kao, C., & Hwang, G. (2020). Facilitating students' critical thinking and decision making performances: A Flipped classroom for neonatal health care training. *Educational Technology & Society*, 23(2), 32–46.
- Chen, K., Monrouxe, L., Lu, Y., Jenq, C., Chang, Y., Chang, Y., & Chai, P. Y. (2018). Academic outcomes of flipped classroom learning: a meta-analysis. *Medical Education*, 52, 910-924.
- Chen, L., Ritzhaupt, A. D., & Antonenko, P. (2019). Effects of the flipped classroom instructional strategy on students' learning outcomes: A Meta-analysis. *Educational Technology Research and Development*, 67, 793-824.
- Chen, M. A. (2019). A Reflective thinking-promoting approach to enhancing graduate students' flipped learning engagement, participation behaviors, reflective thinking and project learning outcomes. *British Journal of Educational Technology*, 50(5), 2288-2307.
- Chen, P. & Hwang, G. (2019). An IRS-facilitated collective issue-quest approach to enhancing students' learning achievement, self-regulation and collective efficacy in flipped classrooms. *British Journal of Educational Technology*, 50(4), 1996-2013.
- Cheng, S., Hwang, G., & Lai, C. (2020). Critical research advancements of flipped learning: A Review of the top 100 highly cited papers. *Interactive Learning Environments*, doi:10.1080/10494820.2020.1765395
- Cheung, C. & Chen, Y. (2020). Implementing the flipped classroom approach in primary English classrooms in China. *Education and Information Technologies*, 25, 1217-1235.
- Chyr, W.-L., Shen, P.-D., Chiang, Y.-C., Lin, J.-B., & Tsia, C.-W. (2017). Exploring the effects of online academic helpseeking and flipped learning on improving students' learning. *Educational Technology & Society*, 20(3), 11–23.
- Davis, C. (2013). Flipped or inverted learning: Strategies for course design. In E. G. Smyth & J. X. Volker (Eds.), *Enhancing instruction with visual media: Utilizing video and lecture capture* (pp. 241-265). Pennsylvania, USA: IGI Global.
- Djamàa, S. (2020). Lecture in the living room, homework in the classroom: The Effects of flipped instruction on graduate EFL students' exam performance. *Computers in the Schools*, 37(3), 141-167.
- Elian, S. A. & Hamaidi, D. A. (2018). The Effect of using flipped classroom strategy on the academic achievement of fourth grade students in Jordan. *International journal of Emerging Technologies in Learning*, 13(2), 110-125.
- Ferriz, A., Sebastián, S., & García, S. (2017). Clase invertida como elemento innovador en Educación Física. Efectos sobre la motivación y la adquisición de aprendizajes en Primaria y Bachillerato [Flipped Classroom as an innovative element in Physical Education: effects on the motivation and the acquisition of learning in Primary and Baccalaureate]. In R. Roig-Vila (Ed.), *Investigación en docencia universitaria. Diseñando el futuro a partir de la innovación educativa* (pp. 211-222). Madrid, Spain: Octaedro.
- Frederick, S. (2002). The Global digital divide: Focusing on children. *Hastings Communications and Entertainment Law Journal*, 24(4), 477-504.
- Galindo-Domínguez, H. (2018). Un meta-análisis de la metodología Flipped Classroom en el aula de Educación primaria [A meta-analysis about flipped classroom methodology in primary education classroom]. *EDUTEC, Revista Electrónica de Tecnología Educativa*, 63, 73-85.
- Galindo-Domínguez, H. (2019a). Efectividad de la metodología Flipped Classroom en la mejora del clima social en el aula de Educación Primaria [Effectiveness of the Flipped Classroom methodology in improving the social climate in the primary education classroom]. *Aloma*, 37(2), 63-69.
- Galindo-Domínguez, H. (2019b). Influence of flipped classroom methodology on the self-concept of primary education students. *Aloma*, 37(2), 35-42.
- Galindo-Domínguez, H. y Bezanilla, M. J. (2019). A Systematic review of flipped classroom methodology at university level in Spain. *Innoeduca. International Journal of Technology and Educational Innovation*, 5(1), 81-90.
- Gillete, C., Rudolph, M., Kimble, C., Rockich-Winston, N., Smith, L., & Broedel-Zaugg, K. (2018). A Meta-analysis of outcomes comparing flipped classroom and lecture. *American Journal of Pharmaceutical Education*, 82(5), 433-440.
- Gómez-García, J., Sellés, S., & Ferriz-Valero, A. (2019). Flipped Classroom Como Propuesta en la Mejora del Rendimiento Académico y Motivación del Alumnado en Educación Física [Flipped Classroom as a Proposal to Improve the Academic Performance and Motivation of the Pupil in Physical Education]. *Revista Kronos*, 18(2), 1-12.
- Herreid, C. F., Schiller, N. A., Herreid, K. F., & Wright, C. B. (2014). A Chat with the survey monkey: Case studies and the flipped classroom. *Journal of College Science Teaching*, 44, 75–80.
- Hung, H. (2015). Flipping the classroom for English language learners to foster active learning. *Computer Assisted Language Learning*, 28(1), 81-96.

- Jdaitawi, M. (2019). The Effect of flipped classroom strategy on students learning outcomes. *International Journal of Instruction*, 12(3), 665-680.
- Jian, Q. (2019). Effects of digital flipped classroom teaching method integrated cooperative learning model on learning motivation and outcome. *The Electronic Library*, 37(5), 842-859.
- Khan, S. (2008). *An Experimental study to evaluate the effectiveness of cooperative learning versus traditional learning method* [International Islamic University]. Retrieved from <http://pr.hec.gov.pk/jspui/bitstream/123456789/414/1/582S.pdf>
- Kumar, K., Chang, C., & Chang, C. (2016). The Impact of the flipped classroom on mathematics concept learning in high school. *Educational Technology & Society*, 19(3), 134-142.
- Kurt, G. (2017). Implementing the flipped classroom in teacher education: Evidence from Turkey. *Educational Technology & Society*, 20(1), 211-221.
- Låg, T. & Grøm, R. (2019). Does the flipped classroom improve student learning and satisfaction? A Systematic review and meta-analysis. *AERA Open*, 5(3), 1-17.
- Lage, M.J., Platt, G., & Treglia, M. (2000). Inverting the classroom: A Gateway to creating an inclusive learning environment. *The Journal of Economic Education*, 31(1), 30-43.
- Lai, T., Ting, F., Yuch, H. (2020). The Effectiveness of team-based flipped learning on a vocational high school economics classroom. *Interactive Learning Environments*, 28(1), 130-141.
- Lin, C.J. & Hwang, G. J. (2018). A Learning analytics approach to investigating factors affecting EFL students' oral performance in a flipped classroom. *Educational Technology & Society*, 21(2), 205-219.
- Liu, C., Sands-Meyer, S., & Audran, J. (2019). The Effectiveness of the student response system (SRS) in English grammar learning in a flipped English as a foreign language (EFL) class. *Interactive Learning Environments*, 27(8), 1178-1191.
- Lo, C. K. & Hew, K. F. (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(4). doi:10.1186/s41039-016-0044-2
- Mahmoud, I. (2020). The Impact of flipped classroom on developing Arabic speaking skills. *The Asia-Pacific Education Researcher*, 29, 295-306.
- Mattis, K. V. (2015). Flipped classroom versus traditional textbook instruction: Assessing accuracy and mental effort at different levels of mathematical complexity. *Technology, Knowledge and Learning*, 20, 231-248.
- Mckay, J. & Emmison, M. (1995). Using learner-centred learning (LCL) in undergraduate sociology courses. *ANZ Journal of Sociology*, 31(3), 94-103.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS medicine*, 6(7), e1000097.
- Novak, G.M. (2011). Just-in-time teaching. *New Directions for Teaching and Learning*, 2011(128), 63-73.
- 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.
- OECD. (2017). *PISA 2015 Results (Volume III): Students' Well-being*. OECD Publishing.
- Pedró, F. (2005). Comparing traditional and ICT-enriched university teaching methods: Evidence from two empirical studies. *Higher Education in Europe*, 30(3), 399-411.
- Prince, M. (2004). Does active learning work? A Review of the research. *Journal of Engineering Education*, 93(3), 223-231.
- Rutkiene, A. & Tandzegolskiene, I. (2015). Students' attitude towards learning methods for self-sufficiency development in higher education. In V. Lubkina & S. Usca (Eds), *Society. Integration. Education. Proceedings of the International Scientific Conference* (pp. 348-357). Rēzekne, Latvia: Rēzeknes Augstskola.
- Schmidt, H. G., Cohen, J., Van der Molen, H. T., Splinter, T. W., Bulte, J., Holdrinet, R., Van Rossum, H. J. (2010). Learning more by being taught less: A "time-for-self-study" theory explaining curricular effects on graduation rate and study duration. *Higher Education*, 60, 287-300.
- Scott, G. W., Furnell, J., Murphy, C. M., & Goulder, R. (2015). Teacher and student perceptions of the development of learner autonomy: A Case study in the biological sciences. *Studies in Higher Education*, 40(6), 945-956.
- Skutil, M., Havlíčková, K., Matějčková, R. (2015). Teaching methods in primary education from the teacher's point of view. In I. Önder (Pres.), *ERPA International Congress on Education* (pp. 26). Les Ulis Cedex, France: SHS Web of Conferences.
- Thomas, L., Hockings, C., Ottaway, J., & Jones, R. (2015). *Independent learning: students' perspectives and experiences*. London, UK: Higher Education Academy.

- Thongkoo, K., Panjaburee, P., & Daungcharone, K. (2019). Integrating inquiry learning and knowledge management into a flipped classroom to improve students' web programming performance in higher education. *Knowledge Management & E-Learning*, 11(3), 304-324.
- Uddin, F. & Khan, N. (2018). Comparing traditional teaching method and experiential teaching method using experimental research. *Journal of Education and Educational Development*, 5(2), 276-288.
- UNICEF (2017). *Children in a digital world*. New York, NY: UNICEF publications.
- Uzunboylu, H. & Karagözlü, D. (2017). The Emerging trend of the flipped classroom: A Content analysis of published articles between 2010 and 2015. *Revista de Educación a Distancia*, 54, 1-13.
- Wasserman, N. H., Quint, C., Norris, S. A., & Carr, T. (2017). Exploring flipped classroom instruction in calculus III. *International Journal of Science and Mathematics Education*, 15, 545-568.
- Wei, X., Cheng, L., Chen, N., Yang, X., Liu, Y., Dong, Y., Zhai, X. & Chatterjee, K. (2020). Effect of the flipped classroom on the mathematics performance of middle school students. *Educational Technology Research and Development*, 68, 1461-1484.

## Appendix: List of references from the systematic review

- Alsancak, D. & Özdemir, S. (2018). The Effect of a flipped classroom model on academic achievement, self-directed learning readiness, motivation and retention. *Malaysian Online Journal of Educational Technology*, 6(1), 76-91.
- Aşıksoy, G. (2018). The Effects of the gamified flipped classroom environment (GFCE) on students' motivation, learning achievements and perception in a physics course. *Quality & Quantity*, 52, 129-145.
- Ayçiçek, B. & Yanpar, T. (2018). The Effect of flipped classroom model on students' classroom engagement in teaching English. *International journal of Instruction*, 11(2), 385-398.
- Beth, A., Hedges, A., Benedict, N. J. & Donihi, A. (2016). Combination of a flipped classroom format and a virtual patient case to enhance active learning in a required therapeutics course. *American Journal of Pharmaceutical Education*, 80(10), 1-8.
- Cabi, E. (2018). The Impact of the flipped classroom model on students' academic achievement. *International Review of Research in Open and Distributed Learning*, 19(3), 1-21.
- Chang, B., Chang, C., Hwang, G., & Kuo, F. (2019). A Situation-based flipped classroom to improving nursing staff performance in advanced cardiac life support training course. *Interactive Learning Environments*, 27(8), 1062-1074.
- Chang, C. & Koong, H. (2019). Classroom interaction and learning anxiety in the IRSIntegrated flipped language classrooms. *Asia-Pacific Edu Res*, 28(3), 193-201.
- Chang, C., Kao, C., & Hwang, G. (2020). Facilitating students' critical thinking and decision making performances: A Flipped classroom for neonatal health care training. *Educational Technology & Society*, 23(2), 32-46.
- Chen, M. A. (2019). A Reflective thinking-promoting approach to enhancing graduate students' flipped learning engagement, participation behaviors, reflective thinking and project learning outcomes. *British Journal of Educational Technology*, 50(5), 2288-2307.
- Chen, P. & Hwang, G. (2019). An IRS-facilitated collective issue-quest approach to enhancing students' learning achievement, self-regulation and collective efficacy in flipped classrooms. *British Journal of Educational Technology*, 50(4), 1996-2013.
- Cheung, C. & Chen, Y. (2020). Implementing the flipped classroom approach in primary English classrooms in China. *Education and Information Technologies*, 25, 1217-1235.
- Chu, T., Wang, J., Monrouxe, L., Sung, Y., Kuo, C., Ho, L., Lin, Y. (2019). The Effects of the flipped classroom in teaching evidence based nursing: A Quasi-experimental study. *Plos ONE*, 14(1), 1-12.
- Chyr, W.-L., Shen, P.-D., Chiang, Y.-C., Lin, J.-B., & Tsia, C.-W. (2017). Exploring the effects of online academic helpseeking and flipped learning on improving students' learning. *Educational Technology & Society*, 20(3), 11-23.
- Craft, E. & Linask, M. (2020). Learning effects of the flipped classroom in a principles of microeconomics course. *The Journal of Economic Education*, 51(1), 1-18.
- Djamàa, S. (2020). Lecture in the living room, homework in the classroom: The Effects of flipped instruction on graduate EFL students' exam performance. *Computers in the Schools*, 37(3), 141-167.
- Elian, S. A. & Hamaidi, D. A. (2018). The Effect of using flipped classroom strategy on the academic achievement of fourth grade students in Jordan. *International journal of Emerging Technologies in Learning*, 13(2), 110-125.

- Eon, J. & Rok, H. (2018). The Impact of flipped learning on cooperative and competitive mindsets. *Sustainability*, 10, 79.
- Eyitayo, E. (2017). The Effects of a flipped classroom model of instruction on students' performance and attitudes towards chemistry. *Journal of Science Education and Technology*, 26, 127-137.
- Fan, J., Tseng, Y., Chao, L., Chen, S., & Jane, S. (2020). Learning outcomes of a flipped classroom teaching approach in an adult-health nursing course: A Quasi-experimental study. *BMC Medical Education*, 20, 317.
- Ferriz, A., Sebastián, S., & García, S. (2017). Clase invertida como elemento innovador en Educación Física. Efectos sobre la motivación y la adquisición de aprendizajes en Primaria y Bachillerato [Flipped Classroom as an innovative element in Physical Education: effects on the motivation and the acquisition of learning in Primary and Baccalaureate]. In R. Roig-Vila (Ed.), *Investigación en docencia universitaria. Diseñando el futuro a partir de la innovación educativa* (pp. 211-222). Madrid, Spain: Octaedro.
- Foldnes, N. (2016). The Flipped classroom and cooperative learning: Evidence from a randomised experiment. *Active Learning in Higher Education*, 17(1), 39-49.
- Galindo-Domínguez, H. (2019a). Efectividad de la metodología Flipped Classroom en la mejora del clima social en el aula de Educación Primaria [Effectiveness of the Flipped Classroom methodology in improving the social climate in the primary education classroom]. *Aloma*, 37(2), 63-69.
- Galindo-Domínguez, H. (2019). Influence of Flipped Classroom methodology on the self-concept of primary education students. *Aloma*, 37(2), 35-42.
- Gómez-García, J., Sellés, S., & Ferriz-Valero, A. (2019). Flipped Classroom Como Propuesta en la Mejora del Rendimiento Académico y Motivación del Alumnado en Educación Física [Flipped Classroom as a Proposal to Improve the Academic Performance and Motivation of the Pupil in Physical Education]. *Revista Kronos*, 18(2), 1-12.
- Haghighi, H., Jafarigohar, M., Khoshsima, H., & Vahdany, F. (2018). Impact of flipped classroom on EFL learners' appropriate use of refusal: Achievement, participation, perception. *Computer Assisted Language Learning*, 32(3), 261-293.
- Hava, K. (2018). The Impact of digital citizenship instruction through flipped classroom model on various variables. *Contemporary educational technology*, 9(4), 390-404.
- Hwang, G. & Lai, C. (2017). Facilitating and bridging out-of-class and in-class learning: An Interactive e-book-based flipped learning approach for math courses. *Educational Technology & Society*, 20(1), 184-197.
- Jdaitawi, M. (2019). The Effect of flipped classroom strategy on students learning outcomes. *International Journal of Instruction*, 12(3), 665-680.
- Jian, Q. (2019). Effects of digital flipped classroom teaching method integrated cooperative learning model on learning motivation and outcome. *The electronic Library*, 37(5), 842-859.
- Jiménez-Millán, A. & Domínguez, J. (2018). Análisis de la eficacia del enfoque flipped learning en la enseñanza de la lengua española en educación primaria [Analysis of the effectiveness of the flipped learning approach in the teaching of Spanish in primary education]. *Didacticae*, 4, 85-107.
- Karabatak, S. & Polat, H. (2020). The Effects of the flipped classroom model designed according to the ARCS motivation strategies on the students' motivation and academic achievement levels. *Education and Information Technologies*, 25, 1475-1495.
- Kazanidis, I., Pellas, N., Fotaris, P., & Tsinakos, A. (2018). Can the flipped classroom model improve students' academic performance and training satisfaction in Higher Education instructional media design courses? *British Journal of Educational Technology*, 50(4), 2014-2027.
- Kırmızı, Ö., & Kömeç, F. (2019). The Impact of the flipped classroom on receptive and productive vocabulary learning. *Journal of Language and Linguistic Studies*, 15(2), 437-449.
- Knežević, L., Županec, V. & Radulović, B. (2020). Flipping the classroom to enhance academic vocabulary learning in an English for Academic Purposes (EAP) Course. *Sage Open*. doi:10.1177/2158244020957052
- Kumar, K., Chang, C., & Chang, C. (2016). The Impact of the flipped classroom on mathematics concept learning in high school. *Educational Technology & Society*, 19(3), 134-142.
- Kurt, G. (2017). Implementing the flipped classroom in teacher education: Evidence from Turkey. *Educational Technology & Society*, 20(1), 211-221.
- Lai, T., Ting, F., Yueh, H. (2020). The Effectiveness of team-based flipped learning on a vocational high school economics classroom. *Interactive Learning Environments*, 28(1), 130-141.
- Lin, C., Hwang, G., Fu, Q. & Chen, J. (2018). A Flipped contextual game-based learning approach to enhancing EFL students' English business writing performance and reflective behaviors. *Educational Technology & Society*, 21(3), 117-131.



- Lin, C.J. & Hwang, G. J. (2018). A Learning analytics approach to investigating factors affecting EFL students' oral performance in a flipped classroom. *Educational Technology & Society*, 21(2), 205-219.
- Liu, C., Sands-Meyer, S., & Audran, J. (2019). The Effectiveness of the student response system (SRS) in English grammar learning in a flipped English as a foreign language (EFL) class. *Interactive Learning Environments*, 27(8), 1178-1191.
- Lo, C. K. & Hew, K. F. (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(4). doi:10.1186/s41039-016-0044-2
- Mahmoud, I. (2020). The Impact of flipped classroom on developing Arabic speaking skills. *The Asia-Pacific Education Researcher*, 29, 295-306.
- Mattis, K. V. (2015). Flipped classroom versus traditional textbook instruction: Assessing accuracy and mental effort at different levels of mathematical complexity. *Technology, Knowledge and Learning*, 20, 231-248.
- Mustapha, H. (2020). Flipped classroom as a supporting plan for Iranian EFL learners' English improvement in super intensive courses. *Theory and Practice in Language Studies*, 10(9), 1101-1105.
- Namaziandost, E. & Çakmak, F. (2020). An Account of EFL learners' self-efficacy and gender in the Flipped Classroom Model. *Education and Information Technologies*, 25, 4041-4055.
- Namaziandost, E. & Shafiee, S. (2018). The Effect of implementing flipped classrooms on iranian junior high school students' reading comprehension. *Theory and Practice in Language Studies*, 8(6), 665-673.
- Penichet-Tomas, A., Jimenez-Olmedo, J. M., Pueco, B., & Carbonell-Martínez, J. A. (2017). Content learning in primary education degree by means of flipped classroom. In L. Góñez., A. López., & I. Candel (Eds.), *EDUlearn 17: 9<sup>th</sup> international conference on education and new lerning technologies* (pp. 3900-3905). Barcelona, Spain: IATED Academy.
- Robert, F. et al. (2017). Randomized controlled study of a remote flipped classroom neuro-otology curriculum. *Frontiers in Neurology*, 8, 349.
- Ruiz, J. L. (2016). El efecto del flipped classroom en la motivación por el aprendizaje del inglés como lengua extranjera en estudiantes de nivel pre-intermedio. *AtoZ: novas práticas em informação e conhecimento*, 5(2), 104-114.
- Schmeisser, C. M. & Medina-Talavera, J. A. (2018). Estudio comparativo entre metodología de aula invertida y metodología tradicional en clases de español, inglés y matemáticas. *MLS-Educational Research*, 2(2), 159-176.
- Sezer, B., & Abay, E. (2018). Looking at the impact of the flipped classroom model in medical education. *Scandinavian Journal of Educational Research*, 63(6), 853-868.
- Sommer, M. & Ritzhaupt, A. D. (2018). Impact of the flipped classroom on learner achievement and satisfaction in an undergraduate technology literacy course. *Journal of Information Technology Education: Research*. *Journal of Information Technology Education: Research*, 17, 159-182.
- Souza, M. J. & Rodrigues, P. (2015). Investigating the effectiveness of the flipped classroom in an introductory programming course. *The New educational review*, 40(2), 129-139.
- Talan, T. & Gulsecen, S. (2019). The Effect of a flipped classroom on students' achievements, academic engagement and satisfaction levels. *Turkish Online Journal of Distance Education*, 20(4), 31-60.
- Thongkoo, K., Panjaburee, P., & Daungcharone, K. (2019). Integrating inquiry learning and knowledge management into a flipped classroom to improve students' web programming performance in higher education. *Knowledge Management & E-Learning*, 11(3), 304-324.
- Wasserman, N. H., Quint, C., Norris, S. A., & Carr, T. (2017). Exploring flipped classroom instruction in calculus III. *International Journal of Science and Mathematics Education*, 15, 545-568.
- Wei, X., Cheng, L., Chen, N., Yang, X., Liu, Y., Dong, Y., Zhai, X. & Chatterjee, K. (2020). Effect of the flipped classroom on the mathematics performance of middle school students. *Educational Technology Research and Development*, 68, 1461-1484.
- Wyk, M. M. (2018). Flipping the economics class in a teacher education course. *Technology, Knowledge and Learning*, 24, 373-399.
- Ye, X., Chang, Y., & Lai, C. (2019). An Interactive problem-posing guiding approach to bridging and facilitating pre- and in-class learning for flipped classrooms. *Interactive Learning Environments*, 27(8), 1075-1092.
- Zainuddin, Z. & Jacqueline, C. (2017). Exploring students' competence, autonomy and relatedness in the flipped classroom pedagogical model. *Journal of Further and Higher Education*, 43(1), 115-126.
- Zheng, M., Chu, C., Jim, Y., Gou, W. (2018). The Mapping of on-line learning to flipped classroom: Small private online course. *Sustainability*, 10(3), 748.

## Interaction Effects of Situational Context on the Acceptance Behaviour and the Conscientiousness Trait towards Intention to Adopt: Educational Technology Experience in Tertiary Accounting Education

Mohamad Ridhuan Mat Dangi<sup>1\*</sup> and Maisarah Mohamed Saat<sup>2</sup>

<sup>1</sup>Faculty of Accountancy, Universiti Teknologi MARA, Selangor Branch, Puncak Alam Campus, Malaysia //

<sup>2</sup>Azman Hashim International Business School, Universiti Teknologi Malaysia, Johor, Malaysia //

ridhuan@uitm.edu.my // maisarahsaat@utm.my

\*Corresponding author

(Submitted November 15, 2020; Revised February 26, 2021; Accepted May 10, 2021)

**ABSTRACT:** The findings of this study reveal that it is unlikely for the interaction effects of situational context, namely educational technology experience (EXP), training frequency (TF), voluntariness (VOL), and class size (CSIZE), to influence accounting educators' intention to adopt educational technology. The original Technology Acceptance Model (TAM), which has been modified numerous times, is still relevant, especially for developing countries since their educational technology penetration is still very low. Conscientiousness trait from the Big Five Personality Model was applied in this study to measure intention as a powerful factor associated with the nature of individuals involved in the accounting profession. Measuring the factors from the individual perspective adds insight into the extant literature since past studies focused on organisational factors and student as the subject. The current study also overcomes the issue of stagnation in the accounting literature, specifically in the field of educational technology. Furthermore, this paper contributes by offering a good indication of using Structural Equation Modelling in the study, especially in the area of accounting and education, and using the most current reporting requirement for information system research.

**Keywords:** Accounting Education, Acceptance Behaviour, Conscientiousness, Educational Technology

### 1. Introduction

The advancement of technology has changed the educators' fundamental activities in teaching-learning, research, scholarship, and service to society (Rana, 2017). Technology is an excellent medium to enhance classroom teaching and learning activities by helping educators to communicate effectively and plan teaching aids and assisting students in self-expression and assertions (Khan, Hasan, & Clement, 2012; Mohd Yusof & Tahir, 2017). Accounting education has also shifted to using educational technology to supplement the pedagogy of 21st century education (Grabinski, Kedzior, & Krasodomska, 2015). The process of teaching and learning accounting subjects requires up-to-date education practices for educators to move from the traditional method of information delivery to contemporary teaching and learning experience (Yisau Abiodun & Tiamiyu, 2012). Therefore, accounting education needs to evolve to fulfil business requirements, prepare students for the market demand and adapt to the changing environment.

A particular concern of past scholars is that accounting educators' role is crucial (O'Connell, Carnegie, Carter et al., 2015), yet they are not using enough technology in the curriculum (Morris, Burnett, Skousen, & Akaaboune, 2015; Burritt & Christ, 2016). Furthermore, employers and industries nowadays are expecting accounting graduates to be equipped with a certain level of accounting skills, a reasonable knowledge of ICT (Ogundana, Ibidunni, & Jinadu, 2015), and deep knowledge of machine learning techniques (ICAEW, 2018) as a new way of thinking and acting of future accountants. The World Economic Forum (2018) predicted that occupation, such as accounting, bookkeeping and payroll clerks are among the top ten declining roles by 2022 due to global change, whereby the role of technology is increasing and changing the role of an accountant. (Morris et al., 2015; Ogundana et al., 2015). Furthermore, the investigation of educational technology research is still low in the Asian and African regions, and the literature is stagnant, especially in the accounting education field (Apostolou, Dorminey, & Hassell, 2020).

Considering the aforementioned concerns, therefore, it becomes the interest of this study to examine the acceptance behaviour and conscientiousness personality traits determinants that may contribute to the intention to use 21st century educational technologies among accounting educators. On top of that, this study also investigates the interaction effects of situational context (e.g., experience in using, training frequency, voluntariness, and class size) with the acceptance behaviour towards the intention to adopt educational technologies in tertiary accounting education.

## 2. Literature review

### 2.1. Educational technology in tertiary accounting education – Experience, issues and recent advancement in teaching practices

Technology and its applications are expanding, and it affects the global economy, leading to radical changes in the accountant's role. Thus, the process of teaching and learning accounting subjects requires a new age of educational practices (Grabinski et al., 2015; Morris et al., 2015; Ogundana et al., 2015). The changing of the accountants' role and accounting process is likely to be affected by how the accounting operates; the influence of technology, storage, processing, retrieval of data, and the process of transactions summarisation (Wells, 2018). Yet, numerous studies on accounting education and technology (e.g., Breedts, 2015; Wong & Wong, 2017; Al-Htaybat, von Alberti-Alhtaybat, & Alhatabat, 2018; Wu, Corr, & Rau, 2019) suggested that a huge gap exists in what is taught by educators in the university and what is being practised in the industry. Besides the audit and database software and general accounting software packages, other common applicable technologies are used in accounting education, such as the internet, e-mail, word processing, presentation, spreadsheets, and data analysis (Ahadiat, 2008; Morris et al., 2015). However, their application is not parallel with the current revolution of advanced technology.

Technology in 21st century teaching and learning is rapidly evolving, with Web 2.0, Web 3.0, virtual reality, e-learning, artificial intelligence, interactive mobile applications, multimedia technology, cloud computing, and other diverse platforms (Watty, McKay, & Ngo, 2016; Al-Htaybat et al., 2018). Therefore, as the key person in spreading technologies, educators need to seize the benefits that come with these innovations by improving their skills and preparing students for a future automated office environment (Nwokike & Eya, 2015; Al-Htaybat et al., 2018). Given the importance of transforming higher education, including the accounting field, Watty, McKay, Ngo et al. (2014) and Adam (2020) proposed ten categories of educational technologies for the accounting teaching and learning activities, which include (1) Learning management systems; (2) Social media or collaborative technologies; (3) Communication technologies; (4) Simulated learning system; (5) Learning style or approach concept; (6) Mobile technologies; (7) Assessment or evaluation technologies; (8) Presentation and learning resource creation tools; (9) Learning objects or resources; and (10) Common accounting tools.

Meanwhile, Yoon (2020) categorised four themes of technologies that can be integrated into accounting education in the digitalisation era, such as Artificial intelligence, Big data, Cloud computing, and Blockchain. These technologies are inevitable; thus, the accounting education field needs to embrace them to prepare future professional accountants with technology and automation knowledge, skills, and abilities. Furthermore, Janvrin and Watson (2017) asserted that the accounting curriculum must be integrated with technology because future accountants will be dealing with a massive volume of business data in the form of a paper-based system and a computer-based system or highly technical enterprise system. This would require proper analytical tools for recording, filtering, summarising, and consolidating the raw data into useful information. Additionally, the application of audit software and knowledge-sharing application using technology in practice indicating a staggering increase of gathering, processing, organising, evaluating, and presenting the financial information (Curtis, Jenkins, Bedard, & Deis, 2009), reporting the business performance, and decision-making process (Pan & Seow, 2016). This is evidenced by the removal of certain manual procedures for presenting the financial information to be aligned with modern business changes (Grabinski et al., 2015; Pincus et al., 2017).

Accordingly, accounting educators are required to respond to this evolution by assimilating with educational technology. It should be endorsed in educational settings to provide students with a new learning experience, given its substantial impact on education and the changes it brings to the pedagogical landscape. Despite the evolution in accounting education, the current scenario suggests that the effort to adopt educational technology is still infancy; both educators and the learners are not utterly familiar (Gaiziuniene & Janiunaite, 2018) with it. Issues, such as unawareness with the changes, lack of interest and knowledge, incompetent (Senik & Broad, 2011; O'Connell et al., 2015; Henriksen et al., 2018; Asonitou, 2020) educators' attitude, resistance to change, and lack of support from the university (Mat Dangi & Mohamed Saat, 2018) are the significant factors leading to the underutilisation of various types of technologies suitable for accounting education.

On top of that, a common dilemma relating to the unsatisfactory level of technology adoption among the accounting professionals, including the academia, is due to the lack of skills, talent leveraging and fails to understand the benefits of instilling technology usage in accounting teaching practices (Malaysian Institute of Accountants, 2018). It is alarming that this situation happens to educational institutions worldwide, especially in developing countries (Abbasi, Tarhini, Hassouna, & Shah, 2015; Khan et al., 2012; Darling-Aduana & Heinrich, 2018), particularly in the 21st century education environment. As the frontline of education, educators'

characteristics and behaviour are crucial elements in scaffolding the efforts to ensure that technology could be successfully integrated into accounting education.

## **2.2. Technology acceptance model (TAM) and the influence of conscientiousness trait**

There have been numerous studies on the adoption, acceptance, intention to use, and usage of information technology in the educational context (Benbasat & Zmud, 1999; Hu, Chau, Sheng & Tam, 1999). However, many researchers are still battling to choose the suitable model or to construct a new model from a number of models (Venkatesh, Morris, Davis, & Davis, 2003), which has been used, altered, and integrated across disciplines, including social sciences, psychology, sociology, education, marketing, information system, and so forth. Weerasinghe and Hindagolla (2017) stated that of all the theories and models (e.g., the theory of reasoned action, unified theory of acceptance and use of technology, diffusion theory, etc.), TAM was the most widely used in many information and technology-related research, and identified as the most robust, parsimonious, and influential model. The technology acceptance model (TAM) was developed by Davis, Bagozzi, and Warshaw (1989) and used extensively by researchers to describe technology acceptance and to determine the reason for an action, whether to accept or reject information technology (Park, 2009).

In this model, there are two direct variables, namely perceived ease of use (E) and perceived usefulness (U) that indicate individuals' intention to utilise an activity while variable of attitude toward using (A) as the mediator predicts the behaviour intention to use (BI) and the intention predicts the behaviour or actual usage. However, the role of attitude towards use as the mediator for PU and PEU is unacceptable since many past studies found it to be a weak intermediary variable to predict the intention and actual usage. Thus, the present study will not remove the attitude construct from TAM, but it will not function as a mediator; instead, it will be one of the direct determinants to measure the accounting educators' intention to adopt educational technology. In a similar vein, Baker, Al-Gahtani, and Hubona (2007), and Altawallbeh, Thiam, Alshourah, and Fong (2015) found that attitude can be a positive determinant that will influence individuals to adopt technology. Additionally, the model application is still relevant in the educational setting. In particular, it can be used to predict the likelihood of new technology adoption in an organisation by groups or individuals (Breedt, 2015). Scherer, Siddiq, and Tondeur (2019) also suggested that TAM is a key model for describing teachers' intention to use technology.

In another perspective, the Big Five personality traits model is one of the most prominent models used in contemporary studies to comprehend the most salient features of personality (Zaidi, Abdul Wajid, Zaidi, Zaidi, & Zaidi, 2013). In particular, early studies provide evidence that personal characteristics have an impact on technology adoption (Xu, Frey, Fleisch, & Ilic, 2016), and it is significantly correlated with people's intention to use the internet, online applications, information sharing, and web browsers (see Tuten & Bosnjak, 2001; Swickert, Hittner, Harris, & Herring, 2002; Amiel & Sargent, 2004; Constantiou, Damsgaard, & Knutsen, 2006; Landers & Lounsbury, 2006). Nonetheless, for this research context, conscientiousness, one of the personality traits, which has been used in past literature, has been shown to be associated with and has an influence on individuals' personality, especially for the type of person working in the accounting profession (Wells, 2003). It should be applicable to study the accounting educators' conscientiousness trait since it is also under the same nature of job background. This would lead to a better understanding as it might imply that the optimal integration of technology into the education field can be achieved.

Moreover, past studies have proved that conscientiousness is also related to behavioural intention and adoption to use hypothetical software technology (Svendson, Johnsen, Almås-Sørensen, & Vittersø, 2013); it positively influences educational performance and work performance in education and learning contexts (Pornsakulvanich, Dumrongsiri, Sajampun et al., 2012) with interesting implications when studying behaviour through intentions (Barnett, Pearson, Pearson & Kellermanns, 2015). Thus, by studying this trait, it is expected that accounting educators with conscientiousness personality trait will be more inclined to have the intention to use technology since these individuals also demonstrate characteristics, such as accountability, dependable, careful, orderly, thoroughness, flexible, and time-saving (Dalpé, Demers, Verner-Filion, & Vallerand, 2019).

## **2.3. Interaction effects of situational context**

Various situational contexts have served as the moderating variables for measuring the interaction effects between exogenous and endogenous constructs. In this study, experience in using educational technology, training frequency, voluntariness, and class size will test the prediction of such variables with educators' acceptance behaviour in tertiary accounting education. Experience in using educational technology, for instance, is used as a moderator since it is associated with individuals' level of knowledge of a new type of system

(Venkatesh et al., 2003). Past literature revealed that the effect of increased experience would impact the acceptance construct (Hartwick & Barki, 1994; Agarwal & Prasad, 1997). Likewise, Hong (2016) mentioned that users' long-term use of technology reflects the users' intention to continue using the technology.

Next is the training frequency, which refers to the efforts of acquiring knowledge and skills required for technology adoption that could improve the technical skills of individuals. As verified in past studies, the efforts of technology training significantly improved the acceptance level of individuals and their intention to adopt technologies (Torkzadeh, Pflughoeft, & Hall, 1999), manage individual perceptions and attitudes about technologies (Marler, Liang, & Dulebohn, 2006); and has a positive influence on the technology acceptance and the intention to use it (Escobar-Rodriguez & Monge-Lozano, 2012). In particular, the study by Mehta (2014) on training elements applied in the e-learning context showed a positive outcome where individuals' technology acceptance is correlated with perceived ease of use and perceived usefulness, and the intention to use the technology. Such training also provides diverse knowledge of the people before and after the training (Smith, 2012). Efficient training programmes provided by the institution can improve educators' level of confidence to easily use the technology, which subsequently develops their intentions to integrate it into their teaching process (Teo, Huang & Hoi, 2018). As a result, effective training will allow the strategy to increase learners' control and engagement (Johnson, List-Ivankovic, Eboh et al., 2010) and decrease the attrition (Salmon 2004) of individuals' acceptance behaviour of the intention to adopt technologies.

In regards to voluntariness, this situational variable is also suggested to have an interaction effect in the context of acceptance behaviour; it was examined in numerous studies on IT acceptance research (Venkatesh, 2000; Venkatesh et al., 2003; Venkatesh & Bala, 2008). It was introduced by Moore and Benbasat (1996) by extending Roger's DOI theory. Past studies also found significant effects of voluntariness variable in mandatory settings, but not in the non-mandatory circumstances (Hartwick & Barki, 1994; Moore & Benbasat, 1996; Agarwal & Prasad, 1997). In a meta-analysis study by Chiu and Ku (2015), it was evidenced that voluntariness context moderates the effects of acceptance behaviour on the intention to use. Such effects were stronger in high-voluntarily settings.

Lastly, the introduction of class size as the moderating factor seems promising in the modernisation of the technology era. A study conducted by Wu, Hsu, and Hwang (2008) found that educational technologies' acceptance and resistance using school factors are unexplored and need to be examined further. Their findings also suggested that educators in small school sizes tended to have a positive attitude towards technology use. The work of Tian, Bian, Han, Gao, and Wang (2017) used class sizes as a moderator in different settings and found that class sizes influenced the academic engagement towards behavioural changes. In this sense, smaller class sizes would inflict less pressure on educators, giving them ample time and opportunities to learn new technologies, offer emotional support and appropriate responses to their students (Beattie & Thiele, 2016; Tian et al., 2017). Thus, this would encourage readiness, acceptance behaviour, and intention to use such technologies in the classroom.

Based on the literature discussed, this study, therefore, postulates the hypotheses (Table 1) for the main effect and interaction effects between the accounting educators' acceptance behaviour and conscientiousness trait with the intention to adopt educational technology.

*Table 1. Hypotheses development for main effect and interaction effects*

Main effect hypotheses	
H1	There is a positive influence of perceived usefulness (ACPU) on the intention to adopt educational technology by the accounting educator
H2	There is a positive influence of perceived ease of use (ACPEU) on the intention to adopt educational technology by the accounting educator
H3	There is a positive influence of attitude towards use (ACAU) on the intention to adopt educational technology by the accounting educator
H4	There is a positive influence of conscientiousness trait (PTCO) on the intention to adopt educational technology by the accounting educator
Interaction Effect Hypotheses for EXP	
H5a	The positive influence between perceived usefulness (ACPU) and the intention to adopt educational technology will be stronger for high experience
H5b	The positive influence between perceived ease of use (ACPEU) and the intention to adopt educational technology will be stronger for high experience
H5c	The positive influence between perceived usefulness (ACAU) and the intention to adopt educational technology will be stronger for high experience

Interaction Effect Hypotheses for TF	
H6a	The positive influence between perceived usefulness (ACPU) and the intention to adopt educational technology will be stronger for frequent training
H6b	The positive influence between perceived usefulness (ACPEU) and the intention to adopt educational technology will be stronger for frequent training
H6c	The positive influence between perceived usefulness (ACAU) and the intention to adopt educational technology will be stronger for frequent training
Interaction Effect Hypotheses for VOL	
H7a	The positive influence between perceived usefulness (ACPU) and the intention to adopt educational technology will be stronger for mandatory
H7b	The positive influence between perceived usefulness (ACPEU) and the intention to adopt educational technology will be stronger for mandatory
H7c	The positive influence between perceived usefulness (ACAU) and the intention to adopt educational technology will be stronger for mandatory
Interaction Effect Hypotheses for CSIZE	
H8a	The positive influence between perceived usefulness (ACPU) and the intention to adopt educational technology will be stronger for small class size
H8b	The positive influence between perceived usefulness (ACPEU) and the intention to adopt educational technology will be stronger for small class size
H8c	The positive influence between perceived usefulness (ACAU) and the intention to adopt educational technology will be stronger for small class size
<i>Note.</i> ACPU – Acceptance Behaviour of Perceived Usefulness; ACPEU - Acceptance Behaviour of Perceived Ease of Use; ACAU - Acceptance Behaviour of Attitude towards Use; PTCO – Personality Trait of Conscientiousness; EXP – Experience; TF – Training Frequency; VOL – Voluntariness; CSIZE – Class Size.	

### 3. Methodology of the study

The convenience sampling and questionnaire survey methods were administered on 275 accounting educators from 12 public universities in Malaysia, offering bachelor's degree programmes in the accounting discipline. About 195 completed responses were received within five months of distribution. The public universities in Malaysia are among the high ranked in the QS World University Ranking, and the number of accounting graduates produced is prominent compared to the private university (Abd Jalil, 2018). The survey questionnaire provided brief information about the definition of intention to adopt and the definition of 21st century educational technology adoption. Since there are limited information pertaining to the technology adoption profile among accounting educators, this study refers 10 categories of educational technologies as outlined by Watty et al. (2014) and Adam (2020), suitably for the 21st century accounting education landscape (see Appendix). The respondents may reflect themselves with any list of educational technologies from the categories they are practicing in the accounting classroom.

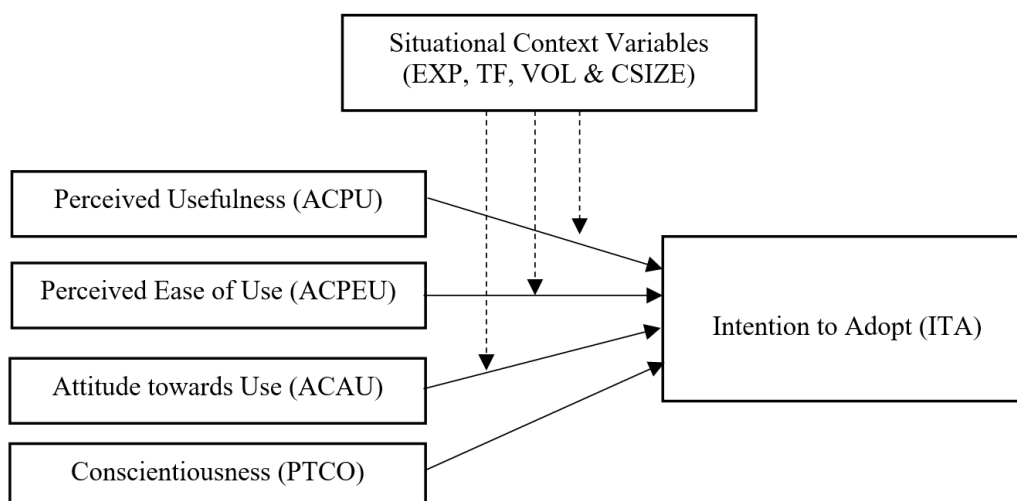


Figure 1. Framework of the study

The items of the survey are segregated into Section A for the demographic profile, Section B with 15 items on the intention to adopt measurements (ITA1–ITA15) with minor modification to suit with the context of the study, 21 adapted items in Section C, assessing the acceptance behaviour of perceived usefulness (ACPU1–ACPU7), perceived ease of use (ACPEU1–ACPEU7), attitude towards use (ACAU1–ACAU2), and conscientiousness trait (PTCO1–PTCO6). The items were assessed using the five-point Likert Scale, ranging from 1 = “Strongly Disagree” to 5 = “Strongly Agree”. Meanwhile, Section C listed four items of situational variables, which were then decoded into categorical variables for interaction effect analysis. Sources for the study were extracted from previous literature of reputable indexed publications authored by Gholami, Abdekhoda, and Gavani (2018), Abu Karsh (2018), Sultan, Woods, and Koo (2011), Agarwal and Prasad (1997), and Barnett et al. (2015). Outputs from this study were analysed using the SmartPLS 3.0 software, following the current requirement and rules of thumb for outer and inner measurement models. Furthermore, the framework of the study (Figure 1) is developed, considering the acceptance behaviour and conscientiousness trait towards the intention to adopt, followed by the testing of the interaction effect of various situational context variables.

## 4. Results and discussion of findings

### 4.1. Profile of respondents

The result of the demographic analysis shows the female respondents were the dominant gender, with 146 (74.9%) compared to the male with 49 (25.1%) respondents. The majority of the respondents were between 40 to 49 years old (53.3%), followed by those between 30 to 39 years old (30.2%), while 16.5% were under 50 years old and above. About 66.2% of the respondents possessed Philosophy Doctorate, whereas 32.8% have a Master’s Degree, and 1% have a professional qualification. In terms of current academic positions, more than half of the respondents (59.5%) are senior lecturers, followed by 22.6% associate professors and 13.8% lecturers. Professors and assistant professors shared the same percentage (2.1%). In this study, about 52.3% of the respondents frequently used educational technologies, while the rest mentioned they used them infrequently.

### 4.2. Assessment of reflective measurement

#### 4.2.1. Internal consistency and convergent validity

This study applies a two-stage modelling technique following the steps recommended by Hair, Sarstedt, Ringle, and Gudergan (2018), the current update of SEM-PLS in information system research by Benitez, Henseler, Castillo, and Schuberth (2020), to develop and examine the reflective measurement model for reliability and validity of the items and constructs, and subsequently to engage with the structural model (Hair, Hult, Ringle & Sarstedt, 2017b). Several assessments have been performed following the rules of thumb, such as internal consistency reliability, convergent validity, and discriminant validity to evaluate the model’s results (Henseler, Ringle, & Sinkovics, 2009; Chin, 2010; Roldán & Sánchez-Franco, 2012; Hair et al., 2017b). The results depicted in Table 2 also include the factor loading estimates of this study. The ranges are from 0.573 to 0.906 and significant at a 1% level, suggesting the measures’ reliability. For this study, all possible outer and inner paths were drawn, and output from the reflective measurement analysis was presented in diagrams and tabulated accordingly.

Table 2. Results for the measurement model

Construct	Indicator	Outer Loadings	Outer Weights	Cronbach Alpha $\alpha$	Dillon–Goldstein’s $\rho$	Dijkstra–Henseler’s $\rho_A$	AVE
Intention to Adopt	ITA2	0.671***	0.127***				
	ITA6	0.646***	0.104***				
	ITA7	0.677***	0.132***				
	ITA8	0.765***	0.150***				
	ITA9	0.771***	0.136***				
	ITA10	0.805***	0.122***	0.906	0.921	0.904	0.52
	ITA11	0.725***	0.113***				
	ITA12	0.700***	0.109***				
	ITA13	0.771***	0.143***				
	ITA14	0.713***	0.145***				
	ITA15	0.638***	0.108***				

Acceptance	ACPU1	0.836***	0.184***				
Behaviour -	ACPU2	0.886***	0.187***				
ACPU	ACPU3	0.849***	0.176***				
	ACPU4	0.816***	0.143***	0.927	0.941	0.927	0.70
	ACPU5	0.816***	0.176***				
	ACPU6	0.884***	0.181***				
	ACPU7	0.751***	0.147***				
Acceptance	ACPEU1	0.817***	0.270***				
Behaviour -	ACPEU2	0.745***	0.216***				
ACPEU	ACPEU3	0.767***	0.141***				
	ACPEU4	0.822***	0.155***	0.875	0.902	0.865	0.57
	ACPEU5	0.790***	0.187***				
	ACPEU6	0.751***	0.210***				
	ACPEU7	0.573***	0.136***				
Acceptance	ACAU1	0.747***	0.163***				
Behaviour -	ACAU2	0.859***	0.167***				
ACAU	ACAU3	0.906***	0.189***				
	ACAU4	0.819***	0.206***	0.927	0.941	0.926	0.70
	ACAU5	0.805***	0.154***				
	ACAU6	0.834***	0.156***				
	ACAU7	0.865***	0.163***				
Personality	PTCO1	0.720***	0.255***				
Trait -	PTCO2	0.727***	0.175***				
PTCO	PTCO3	0.679***	0.213***	0.82	0.87	0.82	0.53
	PTCO4	0.816***	0.278***				
	PTCO5	0.760***	0.232***				
	PTCO6	0.666***	0.211***				
Situational	EXP						
Variable	TF	1.000	1.000	1.000	1.000	1.000	1.000
	VOL						
	CSIZE						

*Note 1.* ACPU – Acceptance Behaviour of Perceived Usefulness; ACPEU - Acceptance Behaviour of Perceived Ease of Use; ACAU - Acceptance Behaviour of Attitude towards Use; PTCO – Personality Trait of Conscientiousness; EXP – Experience; TF – Training Frequency; VOL – Voluntariness; CSIZE – Class Size.

*Note 2.* Situational variables have been decoded into “0” and “1” as the categorical variable.

*Note 3.* Loading indicators are significance when \*\*\*  $p < 0.001$ , (one-tailed test).

In order to achieve the uni-dimensionality of the constructs to ensure all indicators have equal factor scores loaded, the indicator loadings must be above 0.708, indicating that 50% or more of the variance in the observed variables were explained (Hair et al., 2017b). However, for the threshold loadings' value above 0.4, 0.5, 0.6, or 0.7, the indicators will be retained (Wülferth, 2013); if the loadings are below 0.4, then the reflective indicator must be removed from the model (Hulland, 1999; Avkiran & Ringle, 2018). Based on the measurement model in Figure 2, it can be concluded that the majority of the indicator loadings are above 0.5 since the AVE achieved the required minimum threshold of 0.50. Four indicators (ITA1, ITA3, ITA4, and ITA5) were removed one at a time from the lowest loadings, which contributed to the endogenous construct's AVE value of below 0.50. The removal of the items from the model involved only 10% of the whole measurement; thus, it can be assumed that it is a credible instrument design (Hair, Black, Babin, & Anderson, 2010; Hair, Babin & Krey, 2017a), especially when the testing is conducted in the Asia region.

Comparatively, all constructs of the model are considered satisfactory and strongly reliable, whereby both reliability scores assessment criterion using Dillon–Goldstein's  $\rho$  and the strict assessment of PLS consistent Algorithm of Dijkstra–Henseler's  $\rho_A$ , were above 0.70 (Nunnally & Bernstein, 1994; Dijkstra & Henseler, 2015; Hair et al., 2018; Benitez et al., 2020). None of the variable scores from the three assessments' criterion exceeded the problematic values of 0.95, which suggests redundancy.



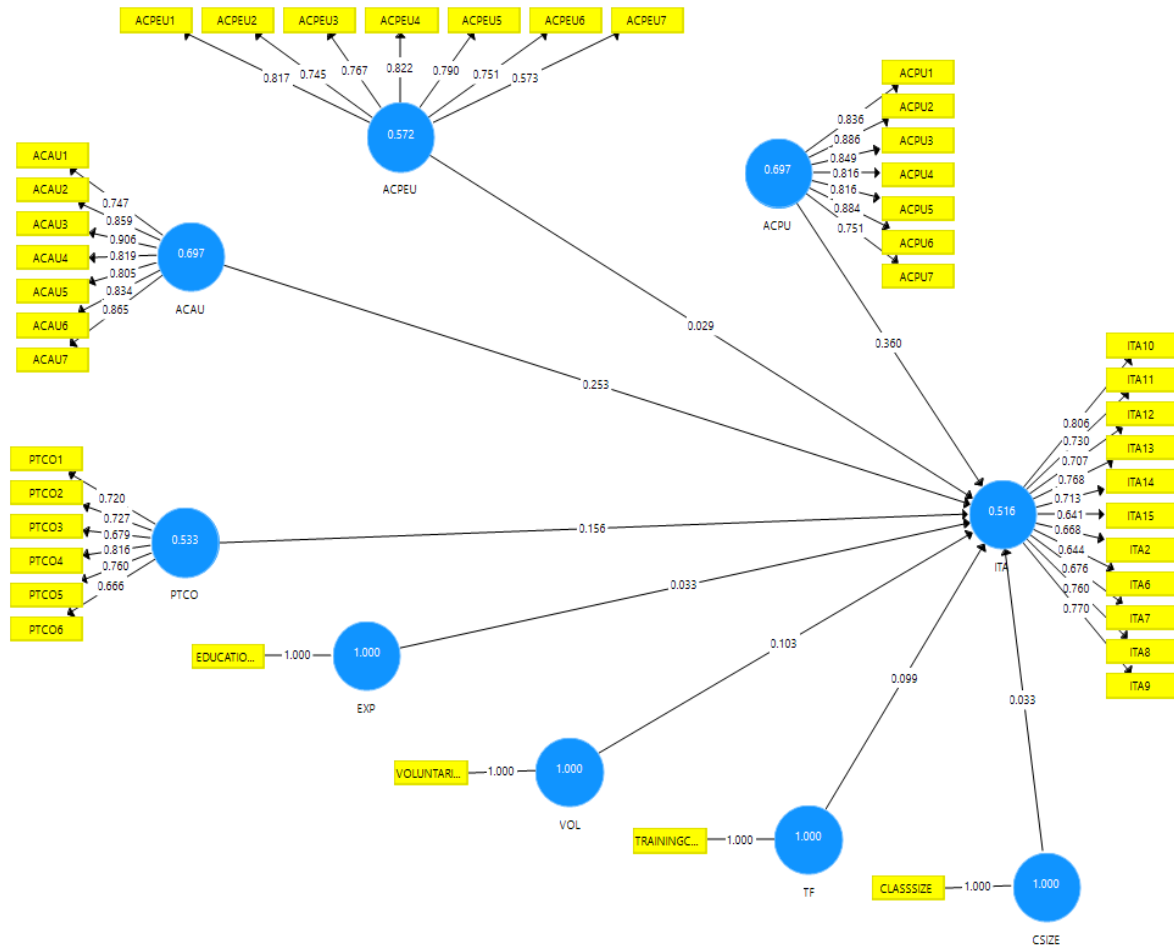


Figure 2. Measurement model of the study

#### 4.2.2. Discriminant validity

The recent discriminant assessment is extended by using the heterotrait-monotrait (HTMT) ratio, as proposed by recent research (Henseler, Ringle, & Sarstedt, 2015; Voorhees, Brady, Calantone, & Ramirez, 2016), particularly in information system research (Benitez et al., 2020). Table 3 illustrates the discriminant validity results of HTMT, which indicate a satisfactory level for all constructs. The HTMT values present a lower value than 0.90 for the lenient threshold and the recommended strict threshold of less than 0.85 (Voorhees et al., 2016; Franke & Sarstedt, 2019). Furthermore, the two-sided of 5% and 95% percentile confidence interval (lower and upper CI) of HTMT does not include the value of 1, indicating that the latent variables are significantly different from 1 on any of the constructs (Henseler et al., 2015); hence, confirming the discriminant validity.

Table 3. HTMT criterion evaluation for discriminant validity

	ITA	ACPU	ACPEU	ACAU	PTCO	EXP	TF	VOL	CSIZE
ITA									
	0.674								
	CI. <sup>95</sup>								
	(0.569,								
	0.757)								
ACPU									
	0.564	0.689							
	CI. <sup>95</sup>	CI. <sup>95</sup>							
	(0.450,	(0.605,							
	0.663)	0.770)							

ACAU	0.661	0.750	0.783				
	CI. <sup>95</sup> (0.550, 0.731)	CI. <sup>95</sup> (0.680, 0.810)	CI. <sup>95</sup> (0.708, 0.841)				
PTCO	0.429	0.265	0.443	0.444			
	CI. <sup>95</sup> (0.306, 0.544)	CI. <sup>95</sup> (0.162, 0.392)	CI. <sup>95</sup> (0.315, 0.548)	CI. <sup>95</sup> (0.335, 0.556)			
EXP	0.219	0.144	0.110	0.168	0.370		
	CI. <sup>95</sup> (0.080, 0.324)	CI. <sup>95</sup> (0.054, 0.244)	CI. <sup>95</sup> (0.053, 0.149)	CI. <sup>95</sup> (0.054, 0.244)	CI. <sup>95</sup> (0.254, 0.490)		
TF	0.383	0.327	0.429	0.380	0.213	0.116	
	CI. <sup>95</sup> (0.276, 0.471)	CI. <sup>95</sup> (0.232, 0.421)	CI. <sup>95</sup> (0.318, 0.521)	CI. <sup>95</sup> (0.269, 0.466)	CI. <sup>95</sup> (0.086, 0.325)	CI. <sup>95</sup> (0.020, 0.190)	
VOL	0.116	0.055	0.111	0.062	0.078	0.108	0.134
	CI. <sup>95</sup> (0.061, 0.176)	CI. <sup>95</sup> (0.032, 0.063)	CI. <sup>95</sup> (0.061, 0.167)	CI. <sup>95</sup> (0.020, 0.115)	CI. <sup>95</sup> (0.031, 0.097)	CI. <sup>95</sup> (0.014, 0.222)	CI. <sup>95</sup> (0.029, 0.244)
CSIZE	0.050	0.093	0.062	0.068	0.081	0.153	0.120
	CI. <sup>95</sup> (0.029, 0.055)	CI. <sup>95</sup> (0.040, 0.173)	CI. <sup>95</sup> (0.025, 0.107)	CI. <sup>95</sup> (0.026, 0.129)	CI. <sup>95</sup> (0.025, 0.127)	CI. <sup>95</sup> (0.019, 0.335)	CI. <sup>95</sup> (0.083, 0.167)

*Note.* ACPU – Acceptance Behaviour of Perceived Usefulness; ACPEU - Acceptance Behaviour of Perceived Ease of Use; ACAU - Acceptance Behaviour of Attitude towards Use; PTCO – Personality Trait of Conscientiousness; EXP – Experience; TF – Training Frequency; VOL – Voluntariness; CSIZE – Class Size.

### 4.3. Assessment of the structural model

#### 4.3.1. Evaluation of path coefficients, significance levels and their effect sizes

Several standard assessment criteria have been applied to assess the structural model, including the coefficient of determination ( $R^2$ ), the blindfolding-based cross-validated redundancy, measuring the  $Q^2$ , and also to test the statistical and relevance of the path coefficients (Hair, Risher, Sarstedt, & Ringle, 2019). Table 4 explains the results' direct effect result of the exogenous and endogenous constructs, as well as the interaction effects. Three hypotheses were supported (H1, H3 and H4), whereby the  $p$ -value  $< .001$  and positively influenced the main effect of endogenous constructs. The coefficients of ACPU ( $\beta_1 = 0.376$ ,  $t = 4.021$ ), ACAU ( $\beta_2 = 0.258$ ,  $t = 2.555$ ), and PTCO ( $\beta_3 = 0.178$ ,  $t = 2.967$ ) showed a significant and strong positive influence of ITA, except the effect of ACPEU. Additionally, using the recommended confidence intervals to measure the results' precision, the percentile bootstrap confidence interval for the path coefficient estimate is considered statistically different from zero at a 5% significance level when its  $p$ -value is below 0.05 or when the 95% bootstrap percentile confidence interval constructed around the estimate does not include zero.

Figure 3 illustrates the  $R^2$  result of 0.483 (48.3%), which is considered substantial (Cohen, 1988) and indicates a strong magnitude of the variance in the intention to adopt, explained and predicted by the exogenous constructs. Moreover, this value above the minimum threshold is widely embraced by many recent literature works (e.g., Benitez et al., 2020; Herrador-Alcaide, Hernández-Solís, & Hontoria, 2020) on the adoption of innovation and information system field. Furthermore, the PLS model of the tested paths demonstrates evidence of predictive relevance, with  $Q^2$  of 0.238 indicating the model's index of reconstruction goodness by model and parameter estimations (Andreev, Heart, Maoz & Pliskin, 2009), which measures the extent of the model's prediction success (Urbach & Ahlemann, 2010).

In this study, the interaction effects using the variables of situational context are used to identify the interactions between the exogenous constructs and endogenous construct. Table 4 shows that three hypotheses are supported (H5c, H6b, and H7b), while the other hypotheses did not exhibit interaction effects between the measured constructs. The significant effects also imply that the confidence interval did not straddle to zero, which signifies the meaningful interaction effects. Meanwhile, the effect sizes of the hypotheses ranged from small to medium. The finding is consistent with many studies in the education field, such as Kraft (2020), which mentioned that

the small effect interpreted by Cohen's standards is often large and meaningful and difficult to achieve large effect sizes (Bakker, Cai, English, Kaiser et al., 2019). Meanwhile, a minimum of 0.02 is recommended for practical significance (Franzblau, 1958; Lipsey, 1998), specifically in the education context.

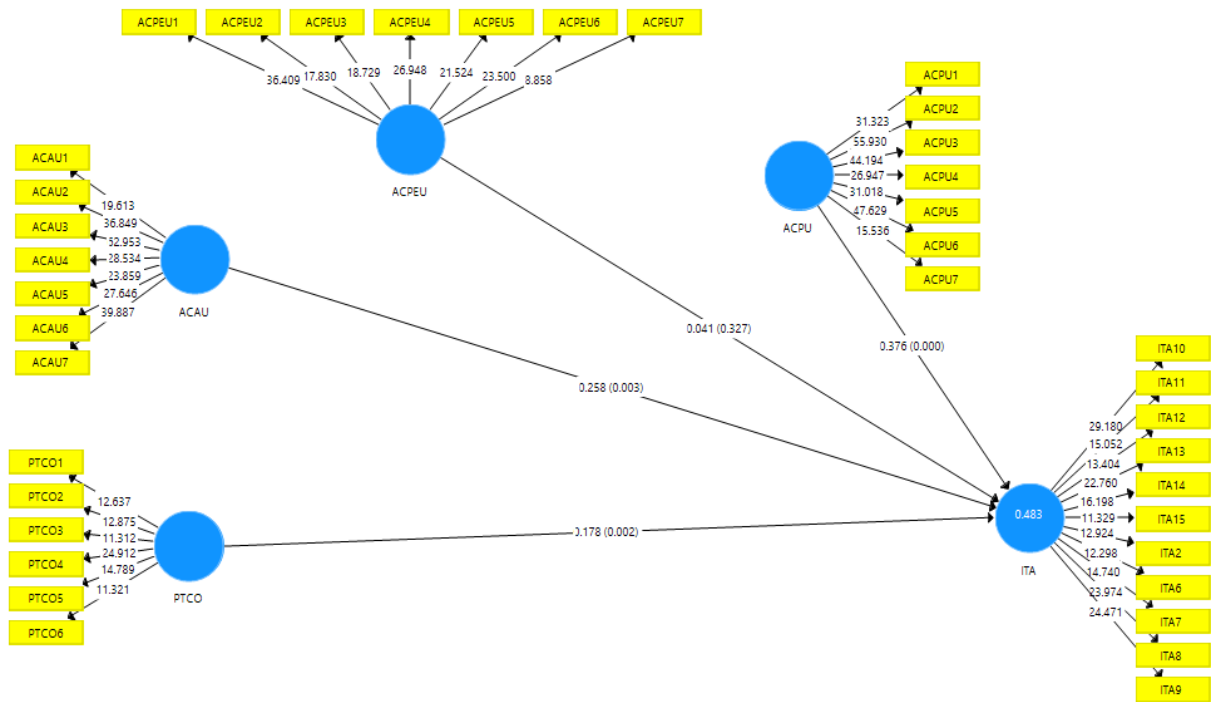


Figure 3. Main effects of the structural model

Table 4. Structural model evaluation

Relationship	Path Coefficient	$f^2$	R <sup>2</sup> Included	R <sup>2</sup> Excluded
<b>Main effects:</b>				
Perceived Usefulness → Intention to Adopt (H1)	0.376*** (4.021) [0.069, 0.408]	0.13		
Perceived Ease of Use → Intention to Adopt (H2)	0.041 (0.478) [-0.122, 0.159]	None		
Attitude towards Use → Intention to Adopt (H3)	0.258*** (2.555) [0.069, 0.408]	0.05		
Conscientiousness → Intention to Adopt (H4)	0.178*** (2.967) [0.071, 0.265]	0.05		
<b>Interaction effects: EXP</b>				
ACPU*EXP → ITA (H5a)	0.411 (0.878) [-0.481, 0.997]	None	0.400	0.392
ACPEU*EXP → ITA (H5b)	0.037 (0.113) [-0.625, 0.470]	None	0.290	0.291
ACAU*EXP → ITA (H5c)	0.516** (1.732) [0.013, 0.971]	0.04	0.423	0.399
<b>Interaction effects: TF</b>				
ACPU*TF → ITA (H6a)	0.149 (1.205) [-0.215, 0.249]	0.02	0.437	0.424
ACPEU*TF → ITA (H6b)	0.145** (2.191) [0.027, 0.240]	0.02	0.327	0.314
ACAU*TF → ITA (H6c)	0.061 (1.009) [-0.071, 0.136]	None	0.415	0.413
<b>Interaction effects: VOL</b>				
ACPU*VOL → ITA (H7a)	-0.323 (1.527) [-0.671, 0.025]	0.03	0.418	0.403
ACPEU*VOL → ITA (H7b)	0.426** (2.276) [0.735, 0.117]	0.04	0.334	0.310
ACAU*VOL → ITA (H7c)	-0.166 (1.042) [-0.429, 0.097]	None	0.419	0.415
<b>Interaction effects: CSIZE</b>				
ACPU*CSIZE → ITA (H8a)	0.131 (0.076) [0.001, 0.319]	None	0.403	0.394
ACPEU*CSIZE → ITA (H8b)	-0.025 (0.060) [-0.180, 0.160]	None	0.288	0.287
ACAU*CSIZE → ITA (H8c)	0.030 (0.034) [-0.237, 0.138]	None	0.394	0.393

Note. *t*-values (one-tailed test) are presented in parentheses. Percentile bootstrap confidence intervals are presented in brackets.

## 5. Discussion on the findings

The findings of the main effects show that perceived usefulness (ACPU), attitude towards use (ACAU), and conscientiousness (PTCO) are significant; thus, they can be predictors to the intention to adopt educational

technology. In line with many other studies, ACPU is one of the strong predictors that influence individuals' intention to adopt educational technology (Al-Marouf & Al-Emran, 2018; Kanwal & Rehman, 2017). Researchers posited that when educators realise the high value of educational technology, they will eventually transform their teaching and learning activities by adopting technology (Akinde & Adetimirin, 2017; McKenney & Visscher, 2019).

Next, attitude (ACAU) variable were also found to be a strong determinant that influences a person's intention to adopt the technology. Therefore, educators with positive attitude towards the use of technology in teaching and learning will potentially be leaning towards implementing or embedding technology in their instructional process (Elkaseh, Wong, & Fung, 2016). This has been highlighted in many studies, which found that a positive attitude towards technology use will result in more efficient use of technology in the teaching and learning process by educators (Guillén-Gámez & Mayorga-Fernández, 2020).

For conscientiousness (PTCO), the significant finding indicates that the accounting educators tend to have similar characters as professional accountants, such as their sensing, thinking, and judging (Bealing, Baker & Russo, 2006), attention to detail, creativity, flexibility, and excellent organisation (Myler, 2021); thus, there is a high probability that educators with high conscientiousness would integrate technology in their instructional activities.

Meanwhile, perceived ease of use (ACPEU) is found to be insignificant. A plausible explanations for this finding could be related to the values and beliefs of accounting educators themselves who not acknowledge the changes of teaching and learning preferences with current needs (Hartman, Townsend & Jackson, 2019), interpreting educational technology as unimportant and not significant to their teaching and learning (Demirbağ & Kılınç, 2018). In this sense, integrating technology in teaching and learning process may be regarded as overwhelming and a burden for accounting educators since it requires much effort to learn and may involve additional costs in terms of financial and time to acquire the skills (Cheung, Wan, & Chan, 2018). In relation to the typical accounting mind-set, accounting educators may assess whether the potential investment in using technology outweighs the cost and guarantee the return (Carlson, 2019). They may use educational technology when it is perceived as useful, meet the learning objectives, and facilitates the instruction process (Akinde & Adetimirin, 2017). In a nutshell, although educational technologies are relatively easy to use and meaningful, the sense of burden, costly and resistance could prevent educators from exploring the opportunities further (Hartman et al., 2019).

On the other hand, of the four situational context variables, three showed interaction effects, namely educational technology experience (EXP), training frequency (TF), and voluntariness (VOL) (excluding the class size [CSIZE]). However, the interaction effects of the three variables affect only one item of TAM. For instance, EXP shows significant interaction effects between ACAU and the intention to adopt technology. Several studies (e.g., Gist, Rosen, & Schwoerer, 1988; Gist, Schwoerer, & Rosen, 1989; Igbaria, Guimaraes & Davis, 1995) corroborate, which suggested that experience could improve individuals' perception and belief about technology use. In particular, when individuals are exposed to technology and have used it for an extended period, it will eventually improve their attitude towards technology use (Hong, 2016).

Correspondingly, training frequency (TF) and voluntariness (VOL) show significant interaction effects between ACPEU and intention to adopt technology, respectively. In view of that, educators could overcome the barriers or anxiety in using technology by getting sufficient training. This is explained by Hu et al. (1999), claimed that training can change individuals' self-efficacy and affect their willingness to adopt technology, including the advanced one. In other words, the number or length of training that educators have will influence their perception of technology's ease of use. Meanwhile, voluntariness (VOL) is the explicit condition that assists in the understanding of individuals' perception of using a specific technology (Venkatesh & Davis, 2000). Thus, educators are likely to use technology, either directly through compliance with mandatory settings or indirectly by recognising the technology's usefulness due to the identification and internationalisation process (Abbasi et al., 2015). The findings of this study are in line with the study by Venkatesh and Davis (2000), which found that individuals will perform a specific behaviour as instructed (in this case, using technology) without prioritising their intentions.

Conversely, the insignificant interaction effects witnessed that perceived usefulness (ACPU) and perceived ease of use (ACPEU) are not moderately influenced by accounting educators' level of experience (EXP) in using technology in their classroom. This is probably related to when individuals are familiarised with technology features and criteria and gain practical experience with them; hence, affecting the perceived ease of use and perceived usefulness would drift away into the background (Tripathi, 2018). Meanwhile, training frequency (TF) and voluntariness (VOL) do not moderately influence both perceived usefulness (ACPU) and attitude (ACAU).

In this sense, ineffective training programmes and poor training content could lead to poor inculcation of positive attitude and difficulties in changing educators' perceptions about the usefulness of technology (Ibrahim, Isa, & Shahbudin, 2016; Akinde & Adetimirin, 2017) in accounting education.

Furthermore, similar to past studies (e.g., Agarwal & Prasad 1997; Chiu & Ku, 2015), voluntariness did not moderately affect attitude as this particular context may be driven by personal interests. Educators either might voluntarily or mandatorily use technology when this effort could meet their personal interests, whether it is favourable or unfavourable (Quazi & Talukder, 2011; Alshmrany & Wilkinson, 2017). Perhaps, the effects between voluntariness and attitude and perceived usefulness could be achieved when accounting educators are surrounded by highly voluntary settings from their colleagues, management, and institutional environment (Fathema, Shannon, & Ross, 2015; Durodolu, 2016; Weerasinghe & Hindagolla, 2017; Opoku & Enu-Kwesi, 2020).

On the other hand, the class size (CSIZE) also did not moderate all the acceptance behaviour constructs (e.g., ACPU, ACPEU and ACAU). This might be explainable as the class size reflects the classroom capacity and may not have a strong influence on strengthening the relationship between accounting educators' acceptance behaviour and their intention to use technology. Educators might think that regardless of the class size, whether big or small, the efforts (e.g., cost, time, skills and knowledge) to prepare themselves with technology would be the same. However, this assumption and perception could be developed and changed when educators have positive attitude, realise the potential benefits of technology in teaching and learning, aware of their need to learn and capture the importance of embracing such technologies in the classroom (Ibrahim et al., 2016).

## **6. Conclusion and future study**

The use of educational technology is widely accepted by educators in many disciplines throughout the world as its benefits are prominently evidenced in the 21st century environment. The merging of e-learning and other educational technology approaches has greatly affected accounting education, whether secondary, tertiary, or professional accounting programmes. Technology affects accounting education in developing the intellectual capital pool by improving teaching quality and inculcating the culture of lifelong learning. This study revealed that TAM and personality traits of conscientiousness could measure individuals' intention towards educational technology. At the same time, characteristics of accountant professionals, especially conscientiousness-related, are reflected. If educational technology is uncomplicated and easy to use but not particularly useful, the intention to adopt is not present and not even considered. The reason being, usefulness implies high return and great benefits for their time, finance, and investment.

Meanwhile, interaction effects' results showed that only experience, training, and voluntariness affect the interaction between certain variables. Moreover, class size did not affect the accounting educators' intention to adopt educational technology in their teaching practices. Several results show the significance and various meaningful indications, especially in the educational context, from small to medium size. Furthermore, this study has its limitations; for example, the sample size of the study might be deemed modest compared with the population. Future studies should consider investigating by using the non-random sampling technique, and also, they could choose individuals who have embraced educational technology for some time or frequently.

In conclusion, the interaction effects' results suggest that the intention to adopt educational technology is derived from the perspectives of individual factors or their attributes. Other factors that might influence their acceptance behaviour will not be affected substantially. Therefore, considering more on individual factors, such as other personality traits that are not commonly associated with individuals in accounting background, would be a meaningful step to generate more findings on the intention to adopt. Future studies may also further explore the accounting educators' characteristics and demographic factors, such as gender, age group, working experience, academic position, income level, and so forth, to understand this technology adoption pattern from an individual perspective. In addition, many studies revolve around students' performance and the impact of using technology. Still, studies on the adoption factors by educators from the academic's perspectives are limited.

## **Acknowledgement**

We would like to appreciate the Faculty of Accountancy, Universiti Teknologi MARA and Azman Hashim International Business School, Universiti Teknologi Malaysia for providing research and materials support.

## References

- Abbasi, M. S., Tarhini, A., Hassouna, M., & Shah, F. (2015). Social, organizational, demography and individuals' technology acceptance behaviour: A Conceptual model. *European Scientific Journal*, 11(9), 48-76.
- Abd Jalil, N. (2018). *Trend pengangguran dalam kalangan graduan di Malaysia* [Unemployment trends among graduates in Malaysia] (Unpublished doctoral dissertation). Universiti Tun Hussein Onn, Malaysia.
- Abu Karsh, S. M. (2018). New Technology Adoption by business faculty in teaching: Analysing faculty technology adoption patterns. *Education Journal*, 7(1), 5-15. doi:10.11648/j.edu.20180701.12
- Adam, B. (2020). *The 101 hottest edtech tools according to education experts* (Updated For 2020). Retrieved from <https://tutorful.co.uk/blog/the-82-hottest-edtech-tools-of-2017-according-to-education-experts>
- Agarwal, R., & Prasad, J. (1997). The Role of innovation characteristics and perceived voluntariness in the acceptance of information technologies. *Decision Sciences*, 28(3), 557-582.
- Ahadiat, N. (2008). Technologies used in accounting education: A Study of frequency of use among faculty. *Journal of Education for Business*, 83(3), 123-134. doi:10.3200/joeb.83.3.123-134
- Akinde, T. A., & Adetimirin, A. A. (2017). Perceived usefulness as a correlate of the extent of information and communications technologies (ICTs) use for teaching by library educators in universities in Nigeria. *International Journal of Library and Information Science*, 9(3), 14-24. doi: 10.5897/IJLIS2016.0739
- Al-Htaybat, K., von Alberti-Alhtaybat, L., & Alhatabat, Z. (2018). Educating digital natives for the future: accounting educators' evaluation of the accounting curriculum. *Accounting Education*, 27(4), 333-357. doi:10.1080/09639284.2018.1437758
- Al-Maroofo, R. A. S., & Al-Emran, M. (2018). Students acceptance of Google Classroom: An Exploratory study using PLS-SEM approach. *International Journal of Emerging Technologies in Learning (iJET)*, 13(6), 112-123. doi:10.3991/ijet.v13i06.8275
- Alshmrany, S., & Wilkinson, B. (2017). Factors influencing the adoption of ICT by teachers in primary schools in Saudi Arabia: Teachers' perspectives of the integration of ICT in primary education. *International Journal of Advanced Computer Science and Applications*, 8(12), 143-156.
- Altawallbeh, M., Thiam, W., Alshourah, S., & Fong, S. F. (2015). The Role of age and gender in the relationship between (attitude, subjective norm and perceived behavioural control) and adoption of e-learning at Jordanian Universities. *Journal of Education and Practice*, 6(15), 44-54.
- Amiel, T., & Sargent, S. L. (2004). Individual differences in Internet usage motives. *Computers in Human Behavior*, 20(6), 711-726. doi: 10.1016/j.chb.2004.09.002
- Andreev, P., Heart, T., Maoz, H., & Pliskin, N. (2009). Validating formative partial least squares (PLS) models: methodological review and empirical illustration. *ICIS 2009 proceedings*, 193. Retrieved from <https://aisel.aisnet.org/icis2009/193>
- Apostolou, B., Dorminey, J. W., & Hassell, J. M. (2020). Accounting education literature review (2019). *Journal of Accounting Education*, 51(C), 100670. doi:10.1016/j.jaccedu.2020.100670
- Asonitou, S. (2020). Technologies to communicate accounting information in the digital era: Is Accounting education following the evolutions? In A. Kavoura, E. Kefallonitis, & P. Theodoridis (Eds.), *Strategic Innovative Marketing and Tourism* (pp. 187-194). 8th ICSIMAT, Northern Aegean, Greece: Springer
- Avkiran, N., & Ringle, C. (2018). Partial least squares structural equation modeling (Vol. 267). In C. C. Price, J. Zhu, & F. S. Hillier (Eds.), *International Series in Operations Research & Management Science*. Cham, Switzerland: Springer.
- Baker, E. W., Al-Gahtani, S. S., & Hubona, G. S. (2007). The Effects of gender and age on new technology implementation in a developing country. *Information Technology & People*, 20(4), 352-375. doi:10.1108/09593840710839798
- Bakker, A., Cai, J., English, L., Kaiser, G., Mesa, V., & Van Dooren, W. (2019). Beyond small, medium, or large: Points of consideration when interpreting effect sizes. *Educational Studies in Mathematics*, 102(1), 1-8. doi:10.1007/s10649-019-09908-4
- Barnett, T., Pearson, A. W., Pearson, R., & Kellermanns, F. W. (2015). Five-factor model personality traits as predictors of perceived and actual usage of technology. *European Journal of Information Systems*, 24(4), 374-390. doi:10.1057/ejis.2014.10
- Bealing, W. E. Jr., Baker, R. L., & Russo, C. J. (2006). Personality: what it takes to be an accountant. *The Accounting Educators' Journal*, 16, 119-128.
- Beattie, I. R., and Thiele, M. (2016). Connecting in class?: College class size and inequality in academic social capital. *The Journal of Higher Education*, 87(3), 332-362. doi: 10.1353/jhe.2016.0017

- Benbasat, I., & Zmud, R. W. (1999). Empirical research in information systems: the practice of relevance. *MIS Quarterly*, 23(1), 3-16. doi: 10.2307/249403
- 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 & Management*, 57(2), 103168. doi:10.1016/j.im.2019.05.003
- Breedt, M. M. (2015). *Aspects influencing Accounting teachers' attitudes towards Computer-Aided Learning* (Unpublished master's thesis). University of Pretoria, South Africa.
- Burritt, R., & Christ, K. (2016). Industry 4.0 and environmental accounting: a new revolution? *Asian Journal of Sustainability and Social Responsibility*, 1(1), 23-38. doi:10.1186/s41180-016-0007-y
- Carlson, T. (2019, May 29). *Here's Why teachers adopt new tech — and why they don't*. Retrieved from <https://www.edsurge.com/news/2019-05-29-here-s-why-teachers-adopt-new-tech-and-why-they-don-t>
- Cheung, G., Wan, K., & Chan, K. (2018). Efficient use of clickers: A Mixed-method inquiry with university teachers. *Education Sciences*, 8(1), 31-46. doi:10.3390/educsci8010031
- Chiu, T. M., & Ku, B. P. (2015). Moderating effects of voluntariness on the actual use of electronic health records for allied health professionals. *JMIR Medical Informatics*, 3(1), e7. doi: 10.2196/medinform.2548
- Chin, W. W. (2010). How to write up and report PLS analyses. In V. V. Esposito, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of partial least squares* (pp. 655-690). Berlin, Heidelberg: Springer.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Constantiou, I. D., Damsgaard, J., & Knutsen, L. (2006). Exploring perceptions and use of mobile services: User differences in an advancing market. *International Journal of Mobile Communications*, 4(3), 231-247. doi:10.1504/IJMC.2006.008940
- Curtis, M. B., Jenkins, J. G., Bedard, J. C., & Deis, D. R. (2009). Auditors' training and proficiency in information systems: A Research synthesis. *Journal of information systems*, 23(1), 79-96. doi:10.2308/jis.2009.23.1.79
- Darling-Aduana, J., & Heinrich, C. J. (2018). The Role of teacher capacity and instructional practice in the integration of educational technology for emergent bilingual students. *Computers & Education*, 126(1), 417-432. doi:10.1016/j.compedu.2018.08.002
- Dalpé, J., Demers, M., Verner-Filion, J., & Vallerand, R. J. (2019). From personality to passion: The Role of the Big Five factors. *Personality and Individual Differences*, 138, 280-285. doi:10.1016/j.paid.2018.10.021
- Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A Comparison of two theoretical models. *Management Science*, 35(8), 982-1003. doi:10.1287/mnsc.35.8.982
- Demirbağ, M., & Kılınç, A. (2018). Preservice teachers' risk perceptions and willingness to use educational technologies: a belief system approach. *Journal of Education and Future*, (14), 15-30. doi:10.30786/jef.379741
- Dijkstra, T. K., & Henseler, J. (2015). Consistent and asymptotically normal PLS estimators for linear structural equations. *Computational statistics & data analysis*, 81(1), 10-23.
- Durodolu, O. O. (2016). *Technology Acceptance Model as a predictor of using information system' to acquire information literacy skills*. (1450). Library Philosophy and Practice (e-journal). Retrieved from <http://digitalcommons.unl.edu/libphilprac/1450>
- Elkaseh, A. M., Wong, K. W., & Fung, C. C. (2016). Perceived ease of use and perceived usefulness of social media for e-learning in Libyan higher education: A Structural equation modeling analysis. *International Journal of Information and Education Technology*, 6(3), 192-199. doi:10.7763/IJiet.2016.V6.683
- Escobar-Rodriguez, T. & Monge-Lozano, P. (2012). The Acceptance of Moodle technology by business administration students. *Computers & Education*, 58(4) 1085-1093. doi:10.1016/j.compedu.2011.11.012
- Fathema, N., Shannon, D., & Ross, M. (2015). Expanding The Technology Acceptance Model (TAM) to examine faculty use of Learning Management Systems (LMSs) in higher education institutions. *MERLOT Journal of Online Learning and Teaching*, 11(2), 210-232.
- Franke, G., & Sarstedt, M. (2019). Heuristics versus statistics in discriminant validity testing: A Comparison of four procedures. *Internet Research*, 29(3), 430-447. doi:10.1108/IntR-12-2017-0515
- Franzblau, A. N., (1958). *A Primer of statistics for non-statisticians*. New York, NY: Harcourt Brace.
- Gaiziuniene, L., & Janiunaite, B. (2018). Adaptation of e-learning tools as innovation: Overcoming barriers using educational factors. In A. O. Mislav, R. Vlasta, & G. Aleksandra (Eds.), *Economic and Social Development: Book of Proceedings* (pp. 403-412). Paris: Varazdin Development and Entrepreneurship Agency.

- Gholami, Z., Abdekhoda, M., & Gavgani, V. Z. (2018). Determinant factors in adopting mobile technology-based services by academic librarians. *DESIDOC Journal of Library & Information Technology*, 38(4). doi:10.14429/djlit.38.4.12676
- Gist, M. E., Rosen, B., & Schwoerer, C. (1988). The Influence of training method and trainee age on the acquisition of computer skills. *Personal Psychology*, 41(2), 255-265. doi:10.1111/j.1744-6570.1988.tb02384.x
- Gist, M. E., Schwoerer, C., & Rosen, B. (1989). Effects of alternative training methods on self-efficacy and performance in computer software training. *Journal of Applied Psychology*, 74(6), 884-891. doi:10.1037/0021-9010.74.6.884
- Grabinski, K., Kedzior, M., & Krasodomska, J. (2015). Blended learning in tertiary accounting education in the CEE region- A Polish perspective. *Journal of Accounting and Management Information Systems*, 14(2), 378-397.
- Guillén-Gámez, F. D., & Mayorga-Fernández, M. J. (2020). Identification of variables that predict teachers' attitudes toward ICT in higher education for teaching and research: A Study with regression. *Sustainability*, 12(4), 1312. doi.org/10.3390/su12041312
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, New Jersey: Prentice-Hall.
- Hair, J. F., Sarstedt, M., Hopkins, L., & Kuppelwieser, V. G. (2014). Partial least squares structural equation modeling (PLS-SEM): An emerging tool in business research. *European Business Review*, 26(2), 106-121. doi:10.1108/EBR-10-2013-0128
- Hair, J. F., Babin, B. J., & Krey, N. (2017a). Covariance-based structural equation modeling in the Journal of Advertising: Review and recommendations. *Journal of Advertising*, 46(1), 163-177.
- Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2017b). *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)* (2nd ed). Thousand Oaks, California: Sage publications Inc.
- Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. *European Business Review*, 31(1), 2-24.
- Hair, J. F., Sarstedt, M., Ringle, C. M., & Gudergan, S. P. (2018). *Advanced issues in partial least squares structural equation modeling (PLS-SEM)*. Thousand Oaks, California: Sage publications Inc.
- Hartman, R. J., Townsend, M. B., & Jackson, M. (2019). Educators' perceptions of technology integration into the classroom: A Descriptive case study. *Journal of Research in Innovative Teaching & Learning*, 12(3), 236-249. doi:10.1108/jrit-03-2019-0044
- Hartwick, J., & Barki, H. (1994). Explaining the role of user participation in information system use. *Management Science*, 40(4), 440-465. doi:10.1287/mnsc.40.4.440
- Henriksen, D., Henderson, M., Creely, E., Ceretkova, S., Černochová, M., Sendova, E., Sointu, E. T., & Tienken, C. H. (2018). Creativity and Technology in Education: An International Perspective. *Technology, Knowledge and Learning*, 23(3), 409-424. doi:10.1007/s10758-018-9380-1
- Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A New criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, 43(1), 115-135. doi:10.1007/s11747-014-0403-8
- Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The Use of partial least squares path modeling in international marketing. *Advances in International Marketing*, 20, 277-319. doi: 10.1108/S1474-7979(2009)0000020014
- Herrador-Alcaide, T. C., Hernández-Solís, M., & Hontoria, J. F. (2020). Online learning tools in the era of m-learning: Utility and attitudes in accounting college students. *Sustainability*, 12(12), 5171. doi:10.3390/su12125171
- Hong, L. C. (2016). *Framework Based Teaching (FBT): Concept paper on the implementation of FBM in the teaching of accountancy courses in the distance learning mode in Malaysia*. Retrieved from <http://library.wou.edu.my/vertical/vf2016-15.pdf>
- Hu, P. J., Chau, P. Y. K., Sheng, O. R. L., & Tam, K. Y. (1999). Examining the technology acceptance model using physician acceptance of telemedicine technology. *Journal of Management Information Systems*, 16(2), 91-112.
- Hulland, J. (1999). Use of partial least squares (PLS) in strategic management research: A Review of four recent studies. *Strategic Management Journal*, 20(2), 195-204.
- Ibrahim, H. I., Isa, A., & Shahbudin, A. S. M. (2016). Organizational support and creativity: The Role of developmental experiences as a moderator. *Procedia Economics and Finance*, 35, 509-514. doi:10.1016/s2212-5671(16)00063-0
- Igbaria, M., Guimaraes, T., & Davis, G. B. (1995). Testing the determinants of microcomputer usage via a structural equation model. *Journal of Management Information Systems*, 11(4), 87-114. doi:10.1080/07421222.1995.11518061
- Janvrin, D. J., & Watson, M. W. (2017). Big data: A New twist to accounting. *Journal of Accounting Education*, 38(C), 3-8. doi:10.1016/j.jaccedu.2016.12.009



- Johnson, N., List-Ivankovic, J., Eboh, W. O., Ireland, J., Adams, D., Mowatt, E., & Martindale, S., (2010). Research and evidence based practice: Using a blended approach to teaching and learning in undergraduate nurse education. *Nurse Education in Practice*, 10(1), 43–47. doi:10.1016/j.nepr.2009.03.012.
- Kanwal, F., & Rehman, M. (2017). Factors affecting e-learning adoption in developing countries—Empirical evidence from Pakistan’s higher education sector. *IEEE Access*, 5, 10968-10978. doi:10.1109/access.2017.2714379
- Khan, M. S. H., Hasan, M., & Clement, C. K. (2012). Barriers to the Introduction of ICT into education in developing countries: The Example of Bangladesh. *International Journal of Instruction*, 5(2), 61-80.
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241-253.
- Landers, R. N., & Lounsbury, J. W. (2006). An Investigation of Big Five and narrow personality traits in relation to internet usage. *Computers in Human Behavior*, 22(2), 283-293. doi:10.1016/j.chb.2004.06.001
- Lipsey, M. W. (1998). Design sensitivity: Statistical power for applied experimental research. In L. Bickman & D. J. Rog (Eds.), *Handbook of Applied Social Research Methods* (pp. 39-68). Thousand Oaks, CA: Sage Publications, Inc.
- Malaysian Institute of Accountants. (2018). *MIA digital technology blueprint: Preparing the Malaysian accountancy profession for the digital world*. Retrived from [https://www.mia.org.my/v2/downloads/resources/publications/2018/07/12/MIA\\_Technology\\_Blueprint\\_Spreads\\_format.pdf](https://www.mia.org.my/v2/downloads/resources/publications/2018/07/12/MIA_Technology_Blueprint_Spreads_format.pdf)
- Marler, J. H., Liang, X., & Dulebohn, J. H. (2006). Training and effective employee information technology use. *Journal of Management*, 32(5), 721-743. doi:10.1177/0149206306292388
- Mat Dangi, M. R., & Mohamed Saat, M. (2018, November 15-16). Accounting educators’ adoption and integration of educational technology in the classroom: A Qualitative study. In *Proceedings of the 32nd International Business Information Management Association Conference, IBIMA 2018-Vision 2020: Sustainable Economic Development and Application of Innovation Management from Regional expansion to Global Growth* (pp. 2254-2264). IBIMA Conference, Seville, Spain.
- McKenney, S., & Visscher, A. J. (2019). Technology for teacher learning and performance. *Technology, Pedagogy and Education*, 28(2), 129-132. doi:10.1080/1475939X.2019.1600859
- Mehta, A. (2014). Technology acceptance of e-learning within a blended vocational course in West Africa. In *Proceedings of the International Conference e-Learning 2014, Multi Conference on Computer Science and Information Systems* (pp. 324–328). Retrieved from <https://files.eric.ed.gov/fulltext/ED557293.pdf>
- Mohd Yusof, M. N., & Tahir, Z. (2017). Kepentingan penggunaan media sosial teknologi maklumat dalam pendidikan IPTA [Importance of Information Technology-Driven Social Media in Public Higher Education Institutions]. *Journal of Social Sciences and Humanities* 12(3), 1-10.
- Moore, G. C., & Benbasat, I. (1996). Integrating diffusion of innovations and the theory of reasoned action models to predict the utilisation of information technology by end-users. In K. Kautz & J. Pries-Heje (Eds.), *Diffusion and adoption of information technology*, (pp. 132-146). Boston, MA: Springer.
- Morris, M., Burnett, R. D., Skousen, C., & Akaaboune, O. (2015). Accounting education and reform: A Focus on pedagogical intervention and its long-term effects. *The Accounting Educators’ Journal*, 25, 67-93.
- Myler, T. (2021). *10 Traits Every Great Accountant Has*. Retrieved from <https://www.accountingweb.com/practice/practice-excellence/10-traits-every-great-accountant-has>
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed). New York, NY: McGraw-Hill.
- Nwokike, F. O., & Eya, G. M. (2015). A Comparative study of the perceptions of accounting educators and accountants on skills required of accounting education graduates in automated Offices. *World Journal of Education*, 5(5), 64-70. doi:10.5430/wje.v5n5p64
- O’Connell, B., Carnegie, G. D., Carter, A. J., de Lange, P., Hancock, P., Helliard, C., & Watty, K. (2015). *Shaping the future of accounting in business education in Australia*. Melbourne, Australia: CPA.
- Ogundana, O. M., Ibidunni, A. S., & Jinadu, O. (2015). ICT Integration in Accounting Education: Evidence from Two Private Higher Institutions in Nigeria. *Acta Universitatis Danubius*, 9(2), 114-126.
- Opoku, M. O., & Enu-Kwesi, F. (2020). Relevance of the technology acceptance model (TAM) in information management research: A Review of selected empirical evidence. *Research Journal of Business and Management*, 7(1), 34-44. doi:10.17261/Pressacademia.2020.1186
- Pan, G., & Seow, P.-S. (2016). Preparing accounting graduates for the digital revolution: A Critical review of information technology competencies and skills development. *Journal of Education for Business*, 91(3), 166-175. doi:10.1080/08832323.2016.1145622
- Park, S. Y. (2009). An Analysis of the Technology Acceptance Model in Understanding University Students’ Behavioral Intention to Use e-Learning. *Journal of Educational Technology & Society*, 12(3), 150-162.

- Pincus, K. V., Stout, D. E., Sorensen, J. E., Stocks, K. D., & Lawson, R. A. (2017). Forces for change in higher education and implications for the accounting academy. *Journal of Accounting Education*, 40, 1-18. doi:10.1016/j.jaccedu.2017.06.001
- Pornsakulvanich, V., Dumrongsiri, N., Sajampun, P., Sornsri, S., John, S. P., Sriyabhand, T., Nuntapanich, C., Chantarawandi, C., Wongweeranonchai, P., & Jiradilok, S. (2012). An Analysis of personality traits and learning styles as predictors of academic performance. *ABAC Journal*, 32(3), 1-19.
- Quazi, A., & Talukder, M. (2011). Demographic determinants of adoption of technological innovation. *Journal of Computer Information Systems*, 51(3), 38-46. doi:10.1080/08874417.2011.11645484
- Rana, K. B. (2017). *Use of educational technologies in teaching and learning activities: Strategies and challenges-A Nepalese case* (Unpublished master's thesis). University of Oslo, Norway.
- Roldán, J. L., & Sánchez-Franco, M. J. (2012). Variance-based structural equation modeling: Guidelines for using partial least squares in information systems research. In M. Mora, O. Gelman, A. L. Steenkamp, & M. Raisinghani (Eds.), *Research methodologies, innovations and philosophies in software systems engineering and information systems* (pp. 193-221). Hershey, PA: IGI Global.
- Salmon G., (2004). *E-moderating: The Key to teaching and learning online* (2nd ed.). London, UK: Routledge Falmer.
- Scherer, R., Siddiq, F., & Tondeur, J. (2019). The Technology acceptance model (TAM): A Meta-analytic structural equation modeling approach to explaining teachers' adoption of digital technology in education. *Computers & Education*, 128, 13-35. doi:10.1016/j.compedu.2018.09.009
- Senik, R., & Broad, M. (2011). Information technology skills development for accounting graduates: Intervening conditions. *International Education Studies*, 4(2), 105-110.
- Smith, A. T. (2012). Middle grades literacy coaching from the coach's perspective. *Research in Middle-Level Education*, 35(5), 1-16. doi:10.1080/19404476.2012.11462088
- Sultan, W. H., Woods, P. C., & Koo, A.-C. (2011). A Constructivist approach for digital learning: Malaysian schools case study. *Journal of Educational Technology & Society*, 14(4), 149-163.
- Svendsen, G. B., Johnsen, J.-A. K., Almås-Sørensen, L., & Vittersø, J. (2013). Personality and technology acceptance: the influence of personality factors on the core constructs of the Technology Acceptance Model. *Behaviour & Information Technology*, 32(4), 323-334. doi:10.1080/0144929X.2011.553740
- Swickert, R. J., Hittner, J. B., Harris, J. L., & Herring, J. A. (2002). Relationships among Internet use, personality, and social support. *Computers in Human Behavior*, 18(4), 437-451. doi:10.1016/S0747-5632(01)00054-1
- Teo, T., Huang, F., & Hoi, C. K. W. (2018). Explicating the influences that explain intention to use technology among English teachers in China. *Interactive Learning Environments*, 26(4), 460-475. doi:10.1080/10494820.2017.1341940
- The Institute of Chartered Accountants in England and Wales (ICAEW). (2018). *Artificial intelligence and the future of accountancy*. Information Technology Faculty. Institute of Chartered Accountants in England and Wales. London. Retrieved from <https://www.icaew.com/technical/technology/artificial-intelligence/artificial-intelligence-the-future-of-accountancy>
- Tian, Y., Bian, Y., Han, P., Gao, F., & Wang, P. (2017). Class collective efficacy and class size as moderators of the relationship between junior middle school students' externalizing behavior and academic engagement: A Multilevel study. *Frontiers in Psychology*, 8, 1219. doi: 10.3389/fpsyg.2017.01219
- Torkzadeh, R., Pflughoeft, K., & Hall, L. (1999). Computer self-efficacy, training effectiveness and user attitudes: an empirical study. *Behaviour & Information Technology*, 18(4), 299-309. doi:10.1080/014492999119039
- Tripathi, S. (2018). Moderating effects of age and experience on the factors influencing the actual usage of cloud computing. *Journal of International Technology and Information Management*, 27(2), 121-158.
- Tuten, T. L., & Bosnjak, M. (2001). Understanding differences in web usage: The Role of need for cognition and the five-factor model of personality. *Social Behavior and Personality: an international journal*, 29(4), 391-398. doi:10.2224/sbp.2001.29.4.391
- Urbach, N., & Ahlemann, F. (2010). Structural equation modeling in information systems research using partial least squares. *Journal of Information Technology Theory and Application*, 11(2), 5-40.
- Venkatesh, V. (2000). Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information Systems Research*, 11(4), 342-365. doi:10.1287/isre.11.4.342.11872
- Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences*, 39(2), 273-315. doi:10.1111/j.1540-5915.2008.00192.x
- Venkatesh, V., & Davis, F. D. (2000). A Theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, 46(2), 186-204. doi:10.1287/mnsc.46.2.186.11926

- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425-478. doi:10.2307/30036540
- Voorhees, C. M., Brady, M. K., Calantone, R., & Ramirez, E. (2016). Discriminant validity testing in marketing: An Analysis, causes for concern, and proposed remedies. *Journal of the Academy of Marketing Science*, 44(1), 119-134.
- Watty, K., McKay, J., Ngo, L., Holt, D., McGuigan, N., Leitch, S., & Kavanagh, M. (2014). Eportfolios in business education. *A National Study of ePortfolio Implementation*. Chartered Accountants Australia and New Zealand. Retrieved from [http://www.buseport.com.au/uploads/2/6/9/9/26997874/monograph\\_ca\\_anz\\_eportfolios\\_in\\_business\\_education\\_pdf.pdf](http://www.buseport.com.au/uploads/2/6/9/9/26997874/monograph_ca_anz_eportfolios_in_business_education_pdf.pdf)
- Watty, K., McKay, J., & Ngo, L. (2016). Innovators or inhibitors? Accounting faculty resistance to new educational technologies in higher education. *Journal of Accounting Education*, 36, 1-15. doi:10.1016/j.jaccedu.2016.03.003
- Weerasinghe, S., & Hindagolla, M. (2017). Technology acceptance model in the domains of LIS and education: A Review of selected literature. *Library Philosophy and Practice (e-journal)*, 1582. Retrieved from <http://digitalcommons.unl.edu/libphilprac/1582>
- Wells, J. T. (2003). Protect small business. *Journal of Accountancy*, 195(3), 26-32.
- Wells, P. K. (2018). How well do our introductory accounting textbooks reflect current accounting practice? *Journal of Accounting Education*, 42, 40-48. doi:10.1016/j.jaccedu.2017.12.003
- Wong, H., & Wong, R. (2017). Students' perceptions of studying accounting information system course. *International Journal of Business Administration*, 8(2), 1-9. doi:10.5430/ijba.v8n2p1
- World Economic Forum. (2018). *Five things to know about the future of jobs*. Retrieved from <https://www.weforum.org/agenda/2018/09/future-of-jobs-2018-things-to-know/>
- Wu, H.-K., Hsu, Y.-S., & Hwang, F.-K. (2008). Factors affecting teachers' adoption of technology in classrooms: Does school size matter? *International Journal of Science and Mathematics Education*, 6(1), 63-85.
- Wu, S. P. W., Corr, J., & Rau, M. A. (2019). How instructors frame students' interactions with educational technologies can enhance or reduce learning with multiple representations. *Computers & Education*, 128, 199-213. doi:10.1016/j.compedu.2018.09.012
- Wülferth, H. (2013). *Managerial discretion and performance in China: Towards Resolving the Discretion Puzzle for Chinese Companies and Multinationals*. Physica, Berlin, Heidelberg. doi:10.1007/978-3-642-35837-1.
- Xu, R., Frey, R. M., Fleisch, E., & Ilic, A. (2016). Understanding the impact of personality traits on mobile app adoption – Insights from a large-scale field study. *Computers in Human Behavior*, 62, 244-256. doi:10.1016/j.chb.2016.04.011
- Yisau Abiodun, B. & Tihamiyu, R. (2012, November 27). The Use of ICT in teaching and learning of accounting education in Nigeria. Paper presented at the 33rd Annual Convention and International Conference of Nigeria Association for Educational Medial and Technology (NAEMT), Emmanuel Alayande College of Education, Oyo State, Nigeria.
- Yoon, S. (2020). A Study on the Transformation of Accounting Based on New Technologies: Evidence from Korea. *Sustainability*, 12(20), 8669. doi:10.3390/su12208669
- Zaidi, N. R., Abdul Wajid, R., Zaidi, F. B., Zaidi, G. B., & Zaidi, M. T. (2013). The big five personality traits and their relationship with work engagement among public sector university teachers of Lahore. *African Journal of Business Management*, 7(15), 1344-1353. doi:10.5897/AJBM12.290

## Appendix: Survey Questionnaire

**Guidelines** for the respondent while answering this survey:

**Definition of Intention to Adopt:** refers to the individual's willingness to perform a given behavior.

**Definition of 21<sup>st</sup> Century Educational Technology Adoption:** the use of any forms of technology-based devices or platforms, tools, approach and resources since 2000s from various areas of knowledge in the design and development of instructional practice for teaching and learning activities.

Based on the definition of the abovementioned terms, the following list of items are examples of the most prevalent educational technology tools, platforms, approach and resources as a reference to reflect your intention to adopt. Perhaps the list is different from your current practice, but it is acceptable as long as it is under the above definition and not limited to the given examples.

*Table 1. Examples of 21<sup>st</sup> Century Educational Technology that can be integrated in Accounting Education*

No.	Categories of Education Technology	Example	Example of Application in Teaching and Learning Environment
1.	Learning Management Systems (LMS)	Moodle; Blackboard; Desire2Learn; iLearn System; MOOCs; i-Folio; Claroline; MyGuru2; Learning Care; Learning Cube; Blackboard; PutraLMS; MyLMS; UFuture.	This software application is often used for documentation, administration, tracking, reporting, delivering educational courses, training programs, or learning and development programmes. It also allows accounting educators to personalise teaching activities to be interactive.
2.	Social Media or Collaborative Technologies	Blogs; Wikis; Twitter; Facebook; Instagram; YouTube; Google Drive; Dropbox; Vimeo; Metacafe;	Social media or collaborative technologies provide powerful means of interaction and communication between the accounting educators and students to discuss any educational-related matters.
3.	Communication	<i>Asynchronous</i> (e.g., Online Discussion Board; e-mail; WhatsApp; WeChat; Telegram)  <i>Synchronous</i> (e.g., Skype; Google Hangout; Adobe Connect; Bloomz; Remind; Sli.do)	The use of communication software and applications provide an alternative way to communicate and help to build a flexible accounting educator-student interaction without space and time boundaries when discussing educational matters.
4.	Simulated Learning Systems – Institutional Customised Development	The Normalised Game; Legends of Learning; Classcraft; SiLAS Solutions; CodaQuest; Animoto, Legends of Learning	Often used to simulate reality, either a system or environment and includes instructional elements to help students to learn, explore, navigate, or obtain information.
5.	Learning Styles or Approach Concept	Gamification; Padlet; Nearpod; Kahoot! Socrative; Blended-Learning; Mobile-Learning; Distance / Online Learning, Peardeck	The application of these approaches provides a different perspective than the traditional teaching practice as it motivates students to engage and participate actively during the teaching and learning activities. Accounting Educators may personalise the content of teaching, create assessments, and have interactive classroom activities.

6.	Mobile Technology	Tablet computer; Smartphones; Mobile Apps (e.g., iOS, Android)	Mobile technology generally used for cellular communication, and also for cooperative learning where accounting educators may provide students with electronic information and educational content, also known as mobile learning or m-learning that assist in the acquisition of knowledge through a variety of mobile devices.
7.	Technology Assessment or Evaluation	Quizlet; Quizlet live; Google classroom; Quizizz; Formative; MOOCs; ZipGrade; Flipgrid; Scan Attendance Manager; Plickers; Kahoot!; Write to Pdf; Google Spreadsheet; Google Form; ClassDojo	The use of these applications helps educators convert to a digital testing environment– tracking and assessing their students’ performance. They also facilitate communication between accounting educators and students and create digital records for students’ growth and development. More importantly, these applications serve as platforms and mediums for teaching, learning, and assessment.
8.	Presentation and Learning Resource Creation Tools	<i>Software</i> (e.g., Adobe Presenter; Voice Recognition Software; Microsoft PowerPoint; Google Slide; Book creator; Adobe Captivate; Screen capture, i.e., Jing, Camtasia; Prezi; Powtoon; Padlet; Nearpod; Google Slides; Canva; PiktoChart; Adobe Acrobat Reader; Showbie; Plotagon Education)  <i>Hardware</i> (e.g., Drawing Tablet, i.e., Wacom; Microphones; In-class Document Reader; Smartphones)	With these applications, accounting educators and students engaged in technological tools and platforms to create presentation and learning resources in a creative, interactive, and enjoyable manner. These applications provide a more engaging way to deliver educational content, accessibility, and better-conveyed presentation.
9.	Learning Objects or Resources	eBooks; Lecture notes or slides; Narrated PowerPoint slides; Podcast, i.e., audio & video; Video lecturers; Instructional videos; Automated video drawings; Flickr; Google Photos; Photobucket; HP Reveal; Aurasma; Google Drives; QR Code Scanner	Learning objects or resources provide tools and the building blocks for the teaching-learning process, prepare the content, learning activities and elements of context for teaching delivery. These applications enable accounting educators to search and access, and reuse objects and resources in learning activities.
10.	Accounting Tools	ATO eTax software; Microsoft ACCESS; Advanced Microsoft Excel; ABSS; Quickbooks; SAS Enterprise Guide; Internet Evidence Finder Forensics, UBS Accounting Software, SQL Accounting Software, ABSS, Mr. Accounting, AutoCount	These accounting tools can be used to manage the process and functions in accounting activities, such as recording and reporting financial information through electronic media and digital platform. Accounting Educators can expose students to these applications in line with the current technological environment.

## SECTION A: RESPONDENT'S PROFILE

Please answer ALL questions by ticking (✓) in the box below the item number that BEST describes your situation.

1. Please specify your **AGE**.

<input type="checkbox"/>	25 – 29 years old
<input type="checkbox"/>	30 – 34 years old
<input type="checkbox"/>	35 – 39 years old
<input type="checkbox"/>	40 – 44 years old
<input type="checkbox"/>	45 – 49 years old
<input type="checkbox"/>	50 years old and above

2. Please specify your **GENDER**.

<input type="checkbox"/>	Male
<input type="checkbox"/>	Female

3. Please specify your highest **EDUCATION LEVEL**.

<input type="checkbox"/>	Philosophy Doctorate (Ph.D.) or DBA
<input type="checkbox"/>	Master Degree
<input type="checkbox"/>	Bachelor Degree
<input type="checkbox"/>	Professional Qualification (ACCA, CIMA, etc.)
<input type="checkbox"/>	Others: _____ ( <i>Please specify</i> )

4. Please specify your **WORKING EXPERIENCE** as an educator.

<input type="checkbox"/>	Below 5 years
<input type="checkbox"/>	6 – 10 years
<input type="checkbox"/>	11 – 15 years
<input type="checkbox"/>	16 – 20 years
<input type="checkbox"/>	21 – 25 years
<input type="checkbox"/>	26 – 30 years
<input type="checkbox"/>	Above 30 years

5. Please specify your **CURRENT ACADEMIC APPOINTMENT**.

<input type="checkbox"/>	Professor
<input type="checkbox"/>	Associate Professor
<input type="checkbox"/>	Assistant Professor
<input type="checkbox"/>	Senior Lecturer
<input type="checkbox"/>	Lecturer
<input type="checkbox"/>	Assistant Lecturer
<input type="checkbox"/>	Tutor
<input type="checkbox"/>	Others: _____ ( <i>Please specify</i> )

6. Average **CLASS SIZE** that you teach normally per semester.

<input type="checkbox"/>	Less than 10 students
<input type="checkbox"/>	10-15 students
<input type="checkbox"/>	16-20 students
<input type="checkbox"/>	21-25 students
<input type="checkbox"/>	26-30 students
<input type="checkbox"/>	More than 30 students
<input type="checkbox"/>	Others: _____ ( <i>Please specify</i> )

7. Please indicate how **OFTEN** you adopt educational technology in your teaching and learning activities.

<input type="checkbox"/>	Not at all
<input type="checkbox"/>	Rarely
<input type="checkbox"/>	Occasionally
<input type="checkbox"/>	Frequently
<input type="checkbox"/>	Almost always
<input type="checkbox"/>	All the time
<input type="checkbox"/>	Others: _____ ( <i>Please specify</i> )

8. Please indicate your **TRAINING LEVEL** of educational technology for teaching and learning purposes.

<input type="checkbox"/>	Not at all
<input type="checkbox"/>	Rarely
<input type="checkbox"/>	Occasionally
<input type="checkbox"/>	Frequently
<input type="checkbox"/>	Almost always

9. Please indicate your **EXPERIENCE** in using educational technology for teaching and learning activities.


<input type="checkbox"/>	Never learned about it formally
<input type="checkbox"/>	Learned, but not used
<input type="checkbox"/>	Learned, and used for at least one semester
<input type="checkbox"/>	Learned, and used it frequently

10. Please indicate your **VOLUNTARINESS** in using educational technology for teaching and learning activities.

<input type="checkbox"/>	Completely free to decide
<input type="checkbox"/>	Self-commitment, drive to adopt
<input type="checkbox"/>	Some mandated, but otherwise free to decide
<input type="checkbox"/>	Mandated in most aspects of teaching

## SECTION B: RESPONDENT'S INTENTION TO ADOPT EDUCATIONAL TECHNOLOGY IN TEACHING AND LEARNING ACTIVITIES

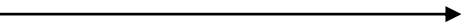
The following questions describe your **INTENTION TO ADOPT educational technology in teaching and learning activities**. Please indicate if you agree or disagree with the following items by using the scale below.

1 = Strongly Disagree  5 = Strongly Agree

	SCALE				
	1	2	3	4	5
1. I will make physical changes to accommodate educational technology in my classroom or computer laboratory.					
2. I will ask my students to use educational technology to enable them to be self-directed learners.					
3. I will use educational technology to record my students' learning activities.					
4. I will share all teaching materials using educational technology with my students.					
5. I will request students to access the teaching materials and resources using educational technology.					
6. I will use educational technology for my teaching management.					
7. I would incorporate educational technology (video, audio, animation) in my teaching and learning activities.					
8. I will conduct the assessment (e.g., quiz, test, simulation test, lab evaluation, project, etc.) using educational technology.					
9. I will instruct my students to use educational technology to complete their assignments and learning activities.					
10. I will motivate the students to communicate and interact using educational technology.					
11. I will ask my students to discuss and collaborate with other students using educational technology platform.					
12. I will use educational technology to encourage my students to share their opinion, response, and idea.					
13. I will perform my students' continuous assessment evaluation using educational technology.					
14. I will evaluate my students' skills acquisition using educational technology.					
15. I will request my students to provide feedback on the teaching and learning using educational technology.					

## SECTION C: RESPONDENT'S ACCEPTANCE BEHAVIOUR AND INTENTION TO ADOPT EDUCATIONAL TECHNOLOGY IN TEACHING AND LEARNING ACTIVITIES

The following questions describe your **ACCEPTANCE BEHAVIOUR and intention to adopt educational technology in teaching and learning activities**. Please indicate if you agree or disagree with the following items by using the scale below.

1 = Strongly Disagree  5 = Strongly Agree

	SCALE				
	1	2	3	4	5
<b>ACPU</b>					
1. I perceive that educational technology enhances my instructive effectiveness in teaching and learning activities.					
2. I perceive that educational technology increases my performance and productivity in teaching and learning activities.					
3. I perceive that educational technology enables me to accomplish tasks in teaching and learning activities more quickly					
4. I perceive that educational technology makes my teaching and learning activities more effective.					
5. I perceive that educational technology gives greater control over my work in teaching and learning activities.					
6. I perceive that educational technology improves the quality of my work in teaching and learning activities.					



7. I perceive that educational technology supports the development of learning outcome in teaching and learning activities.					
<b>ACPEU</b>					
1. I have a clear and understandable interaction with educational technology for teaching and learning activities.					
2. I think the interaction with educational technology in teaching and learning activities is satisfying.					
3. I perceive that learning to operate educational technology and to apply it in teaching and learning activities is not complicated.					
4. I think it is not difficult to remember how to perform tasks using educational technology in teaching and learning activities.					
5. I perceive that interaction with educational technology in teaching and learning activities is flexible.					
6. I think I could become skilful at using technology in teaching and learning activities.					
7. I perceive that interaction with technology in teaching and learning activities does not require much effort.					
<b>ACAU</b>					
1. I think it is fun to use educational technology in teaching and learning activities.					
2. I look forward to the aspects of my job that require me to use educational technology.					
3. I feel passionate about using educational technology for my teaching and learning activities.					
4. I think I am satisfied with using educational technology in teaching and learning activities.					
5. I feel eager when my friends are talking about educational technology.					
6. I am excited when I am working with many types of educational technology in teaching and learning activities.					
7. I am enthusiastic when using educational technology in teaching and learning activities.					

### CONSCIENTIOUSNESS TRAIT AND INTENTION TO ADOPT EDUCATIONAL TECHNOLOGY IN TEACHING AND LEARNING ACTIVITIES.

The following questions describe your **CONSCIENTIOUSNESS TRAIT and intention to adopt educational technology in teaching and learning activities**. Please indicate if you agree or disagree with the following items by using the scale below.

**1 = Strongly Disagree** —————→ **5 = Strongly Agree**

	SCALE				
	1	2	3	4	5
<b>PTCO</b>					
1. I am always prepared					
2. I do not waste my time					
3. I find it is not difficult to get ready to work					
4. I perform a job efficiently					
5. I carry out my plans					
6. I am carefully in my duties					

**Thank you for your participation!**

## Editorial Note: From Conventional AI to Modern AI in Education: Re-examining AI and Analytic Techniques for Teaching and Learning

Haoran Xie<sup>1</sup>, Gwo-Jen Hwang<sup>2\*</sup> and Tak-Lam Wong<sup>3</sup>

<sup>1</sup>Department of Computing and Decision Sciences, Lingnan University, Hong Kong SAR // <sup>2</sup>Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology, Taiwan//

<sup>3</sup>Department of Computing Studies and Information Systems, Douglas College, Canada // hrxie2@gmail.com // gjhwang.academic@gmail.com // tlwong@ieee.org

\*Corresponding author

**ABSTRACT:** With the rapid development and significant successfulness of various deep learning techniques in artificial intelligence (AI) in recent years, the connotation of AI has been transformed from traditional rule-based or statistical learning models to deep learning models. Such a transformation of AI has led to a significant evolution in both academic and industrial fields. To understand the potential impact of AI evolution for future teaching and learning, it is necessary to re-examine the opportunities, research issues, and roles of AI in education as modern AI enables the possibility of playing vital roles in education, which are not only limited to intelligent tutors/tutees but also intelligent learning partners or policy making advisors. Motivated by the recent transformation and trends in AI in education, this special issue, including 12 research articles, aims to launch an in-depth discussion on re-examining AI and analytics techniques in teaching and learning applications.

**Keywords:** Modern AI, AI transformation, Deep neural networks, Analytic techniques, AI in education

### 1. Paradigm shift of AI

There have been various definitions of the term “artificial intelligence (AI)” in the community of computer science. Different from “human intelligence,” AI refers to “computers that mimic cognitive functions that humans associate with the human mind, such as learning and problem-solving” (Russell & Norvig, 2009, p. 2). Russell and Norvig (2009) argued that AI could be defined from the perspective of the intelligent agent, which can perceive the percepts from the external environment and take actions through the effectors to adapt to the environment changes or achieve certain goals. Moreover, Poole and Mackworth (2010, p.1) defined AI as “a system that acts intelligently: What it does is appropriate for its circumstances and its goal, it is flexible to changing environments and changing goals, it learns from experience, and it makes appropriate choices given perceptual limitations and finite computation.”

Although AI is not a new term, the meaning of modern AI has changed compared to conventional AI techniques. (Chen et al., 2020b). Recently, modern AI has tended to refer to the Deep Neural Networks (DNN) based techniques developed in recent years (Yosinski et al., 2014). DNN-based AI and analytic techniques have led to a significant evolution in both academic and industrial fields. With the rapid development of modern AI and analytics techniques such as convolutional neural networks (CNN), generative adversarial networks (GAN), reinforcement learning (RL), and so on, which are based on DNN paradigms, in recent years, there have been a huge number of innovative applications in various domains. For example, long short-term memory (LSTM) techniques have been exploited for predicting stock market prices (Sirignano, & Cont, 2019); CNN techniques have been adopted in surveillance systems, and in self-driving cars (Hu & Ni, 2017; Chen et al., 2017) and RL methods have created some famous AI applications such as Alpha GO (Silver et al., 2016).

### 2. Modern AI in education: Gaps and directions

The research studies about applications of “AI in education” have been conducted for several years. However, the integration of “AI in education” (AIEd) focuses on the use of traditional AI techniques based on rule/statistic-based models to facilitate teaching and learning in education in the past few years. Due to the evolution of AI techniques from rule/statistic-based to DNN-based models in recent years, there has been a limited number of studies on the integration of “modern AI and education” which are based on DNN-based models for teaching and learning. As reported in Chen et al. (2020b), there are only two studies on Modern AI in education (i.e., deep learning in education) among all 45 highly cited AIEd studies in the recent decade. However, the overall trend of AIEd studies has rapidly increased in recent years (Chen et al., 2020a; Hwang et al., 2020). In other words, the potential power of modern AI and analytics applications in education has not been fully exploited or released. The underlying reasons can be divided into two aspects: (1) there is a knowledge gap between AI experts and

educational researchers; and (2) it is quite challenging to integrate the two areas and identify an intersection of valuable applications. To be more specific, AI experts typically do not have knowledge of pedagogical methodologies or in-depth experiences in the classroom, while it is unrealistic to ask educational researchers to be equipped with domain knowledge of modern AI techniques.

From the perspective of education technology, Johnson et al. (2016) published a horizon report which claimed that (i) “Bring Your Own Device” (BYOD) and “Learning Analytics and Adaptive Learning” can be achieved in the near-term (i.e., 1 year or less); (ii) “Augmented and Virtual Reality (AR/VR)” and “Makerspaces” can be achieved in the mid-term (i.e., 2 to 3 years); and (iii) “Affective Computing” and “Robotics” can be achieved in the long-term (i.e., 4 to 5 years). The modern AI applications can facilitate the better adoption of these education technologies in the following aspects:

- Providing modern AI-based learning analytics and adaptive learning. For example, AI-based agents can collect personal information and predict learners’ preferences or learning paths (Xie et al., 2017; Almohammadi et al., 2016; Zou et al., 2020; Wang et al., 2021).
- Facilitating modern AI-based interaction in VR/AR learning environments. For example, AI-based games in VR/AR can better foster learners’ immersion and interaction compared to games without AI (Rahimi & Ahmadi, 2017; Hammedi et al., 2020).
- Supporting affective computing/robotics with highly accurate modern AI models. For example, some deep neural networks can be adopted for analyzing bio-feedback signals such as EEG or brainwaves, which are collected from affective computing devices (Goh et al., 2017; Chen et al., 2021).
- Developing innovative learning applications with modern AI techniques. For example, some recent AI-techniques such as generative adversarial networks (GAN) can create new images, videos, or styles (Mao et al., 2019), which can be employed in drawing learning (Jin et al., 2019; Sorin et al., 2020).

In addition to the above modern AI-enabled applications in education, the transformation from conventional AI to modern AI has led to the reconceptualization of pedagogical innovations. Yang (2021, p. 106) has proposed that precision education, which refers to “identify[ing] at-risk students as early as possible and provid[ing] them with timely intervention through diagnosis, prediction, treatment, and prevention,” is a new challenge for AI in education. Yang et al. (2021) have further developed a conceptual framework by re-organizing precision education as one of the core components of human-centered AI in education, the other components of which are smart learning analytics and smart assessment. By considering the potential applications of modern AI techniques in education, Hwang et al. (2020) have defined a role framework for AIED, which can be categorized as “intelligent tutor,” “intelligent tutee,” “intelligent learning tool/partner,” and “policy-making advisor.”

### 3. The Published papers of this special issue

There were 42 submissions to this special issue. After an initial screening and two rounds of double blinded review, 12 research papers were accepted for publication in this special issue, which can be further divided into five categories. These five categories are (i) AIED systematic review; (ii) modern AI applications; (iii) smart learning environments; (iv) AI-driven interventions; and (v) teaching/learning innovations for AI.

One paper conducted a systematic review of modern AI in education: the paper authored by Fengying Li, Yifeng He, and Qingshui Xue, entitled “Progress, Challenges and Countermeasures of Adaptive Learning: a Systematic Review.”

Four papers have integrated modern AI models such as DNN, LSTM, and BERT for educational applications: the paper authored by Chia-An Lee, Jian-Wei Tzeng, Nen-Fu Huang, and Yu-Sheng Su, entitled “Prediction of Student Performance in Massive Open Online Courses Using Deep Learning System Based on Learning Behaviors”; the paper authored by Albert C. M. Yang, Irene Y. L. Chen, Brendan Flanagan, and Hiroaki Ogata, entitled “Automatic Generation of Cloze Items for Repeated Testing to Improve Reading Comprehension”; the paper authored by Owen H.T. Lu, Anna Y.Q. Huang, Danny C. L. Tsai, and Stephen J.H. Yang, entitled “Expert-Authored and Machine-Generated Short-Answer Questions for Assessing Students’ Learning Performance”; and the paper authored by Changqin Huang, Xuemei Wu, Xizhe Wang, Tao He, Fan Jiang, and Jianhui Yu, entitled “Exploring the relationships between achievement goals, community identification and online collaborative reflection: A deep learning and Bayesian approach.”

Three papers discuss the development of smart learning environments, employing Artificial Intelligence of Things (AIoT) or intelligent agents/robots for improving cognitive and affective factors of learners: the paper

authored by Beyin Chen, Gwo-Haur Hwang, and Shen-Hua Wang, entitled “Gender Differences in Cognitive Load when Applying Game-Based Learning with Intelligent Robots”; the paper authored by Jian-Hua Han, Keith Shubeck, Geng-Hu Shi, Xiang-En Hu, Lei Yang, Li-Jia Wang, Wei Zhao, and Qiang Jiang, entitled “Teachable Agent Improves Affect Regulation: Evidence from Betty’s Brain”; and the paper authored by Chuang-Kai Chiu and Judy C. R. Tseng, entitled “A Bayesian Classification Network-based Learning Status Management System in an Intelligent Classroom.”

Two papers investigate the effects/factors of adopting AI-driven interventions: the paper authored by Youmei Wang, Chenchen Liu, and Yun-fang Tu, entitled “Factors affecting the adoption of AI-based applications in higher education: An analysis of teachers’ perspectives using structural equation modeling”; and the paper authored by Lanqin Zheng, Lu Zhong, Jiayu Niu, Miaolang Long, and Jiayi Zhao, entitled “Effects of Personalized Intervention on Collaborative Knowledge Building, Group Performance, Socially Shared Metacognitive Regulation, and Cognitive Load in Computer-Supported Collaborative Learning.”

Two papers discuss AI teaching and learning innovations: the paper authored by Ching Sing Chai, Pei-Yi Lin, Morris Siu-Yung Jong, Yun Dai, Thomas K.F. Chiu, and Jianjun Qin, entitled “Perceptions of and behavioral intentions towards learning artificial intelligence in primary school students”; and the paper authored by Chun-Hung Lin, Chih-Chang Yu, Po-Kang Shih, and Leon Yufeng Wu, entitled “STEM-based Artificial Intelligence Learning in General Education for Non-Engineering Undergraduate Students.”

## 4. Conclusion

As discussed in the previous sections, the recent breakthrough of modern AI techniques has led a revolution in education. Such a transformation not only involves technical changes in the adoption of modern AI techniques in education, but also reconceptualizes the pedagogical framework in the future. By organizing this special issue, we can clearly foresee that modern AI in education is one of the core research topics in education communities. Furthermore, we can observe the rapid development of the integration of modern AI and education, which have established a mutually driven relationship: the development of modern AI techniques provides a great number of educational applications, while educational innovations have created a critical need for AI-enabled systems.

## Acknowledgement

This study is supported in part by the Ministry of Science and Technology of Taiwan under contract numbers MOST-109-2511-H-011-002-MY3 and MOST-108-2511-H-011-005-MY3.

## References

- Almohammadi, K., Hagra, H., Alghazzawi, D., & Aldabbagh, G. (2016). Users-centric adaptive learning system based on interval type-2 fuzzy logic for massively crowded E-learning platforms. *Journal of Artificial Intelligence and Soft Computing Research*, 6(2), 81-101.
- Chen, X., Ma, H., Wan, J., Li, B., & Xia, T. (2017). Multi-view 3D object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1907-1915). Retrieved from <https://arxiv.org/abs/1611.07759v3>
- Chen, X., Xie, H., & Hwang, G. J. (2020a). A Multi-perspective study on Artificial Intelligence in Education: grants, conferences, journals, software tools, institutions, and researchers. *Computers and Education: Artificial Intelligence*, 1, 100005.
- Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020b). Application and theory gaps during the rise of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 1, 100002.
- Chen, X., Tao, X., Wang, F. L., & Xie, H. (2021). Global research on artificial intelligence-enhanced human electroencephalogram analysis. *Neural Computing and Applications*. doi:10.1007/s00521-020-05588-x.
- Goh, S. K., Abbass, H. A., Tan, K. C., Al-Mamun, A., Wang, C., & Guan, C. (2017). Automatic EEG artifact removal techniques by detecting influential independent components. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 1(4), 270-279.

- Hammedi, S., Essalmi, F., Jemni, M., & Qaffas, A. A. (2020). An investigation of AI in games: educational intelligent games vs non-educational games. In *2020 International Multi-Conference on Organization of Knowledge and Advanced Technologies (OCTA)* (pp. 1-4). IEEE.
- Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 1, 100001.
- Hu, L., & Ni, Q. (2017). IoT-driven automated object detection algorithm for urban surveillance systems in smart cities. *IEEE Internet of Things Journal*, 5(2), 747-754.
- Jin, Y., Li, P., Wang, W., Zhang, S., Lin, D., & Yin, C. (2019). GAN-based pencil drawing learning system for art education on large-scale image datasets with learning analytics. *Interactive Learning Environments*. doi:10.1080/10494820.2019.1636827
- Johnson, L., Becker, S. A., Cummins, M., Estrada, V., Freeman, A., & Hall, C. (2016). *NMC horizon report: 2016 higher education edition*. The New Media Consortium. Retrieved from <https://library.educause.edu/resources/2016/2/2016-horizon-report>
- Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., & Smolley, S. P. (2019). On the effectiveness of least squares generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(12), 2947-2960.
- Poole, D. L., & Mackworth, A. K. (2010). *Artificial Intelligence: Foundations of computational agents*. New York, NY: Cambridge University Press.
- Rahimi, E., & Ahmadi, A. (2017). An AI-based tennis game by application of virtual reality components. In *2017 Iranian Conference on Electrical Engineering (ICEE)* (pp. 2165-2170). doi:10.1109/IranianCEE.2017.7985421
- Russell, S. J., & Norvig, P. (2009). *Artificial Intelligence: A Modern approach* (3rd ed.). Upper Saddle River, New Jersey: Prentice-Hall.
- Sirignano, J., & Cont, R. (2019). Universal features of price formation in financial markets: Perspectives from deep learning. *Quantitative Finance*, 19(9), 1449-1459.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., & Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484-489.
- Sorin, V., Barash, Y., Konen, E., & Klang, E. (2020). Creating artificial images for radiology applications using generative adversarial networks (GANs)—A Systematic review. *Academic Radiology*, 27(8), 1175-1185.
- Wang, J., Xie, H., Wang, F. L., Lee, L. K., & Au, O. T. S. (2021). Top-n personalized recommendation with graph neural networks in MOOCs. *Computers and Education: Artificial Intelligence*, 2, 100010.
- Xie, H., Zou, D., Wang, F. L., Wong, T. L., Rao, Y., & Wang, S. H. (2017). Discover learning path for group users: A Profile-based approach. *Neurocomputing*, 254, 59-70.
- Yang, S. J. H. (2021). Guest editorial: Precision education - A New challenge for AI in education. *Educational Technology & Society*, 24(1), 105-108.
- Yang, S. J., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 2, 100008.
- Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems* (pp. 3320-3328). Retrieved from <https://arxiv.org/abs/1411.1792>
- Zou, D., Wang, M., Xie, H., Cheng, G., Wang, F. L., & Lee, L. K. (2021). A Comparative study on linguistic theories for modeling EFL learners: Facilitating personalized vocabulary learning via task recommendations. *Interactive Learning Environments*, 29(2), 270-282.

## Perceptions of and Behavioral Intentions towards Learning Artificial Intelligence in Primary School Students

Ching Sing Chai<sup>1</sup>, Pei-Yi Lin<sup>2</sup>, Morris Siu-Yung Jong<sup>1\*</sup>, Yun Dai<sup>1</sup>, Thomas K. F. Chiu<sup>1</sup>, and Jianjun Qin<sup>3</sup>

<sup>1</sup>Department of Curriculum and Instruction & Centre for Learning Sciences and Technologies, The Chinese University of Hong Kong, Hong Kong, China // <sup>2</sup>Department of Education, National Kaohsiung Normal University, Taiwan // <sup>3</sup>School of Mechanical-Electronic and Vehicle Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China // cschai@cuhk.edu.hk // pylin@nknku.edu.tw // mjong@cuhk.edu.hk // yundai@cuhk.edu.hk // tchiu@cuhk.edu.hk // qinjianjun@bucea.edu.cn

\*Corresponding author

**ABSTRACT:** Artificial Intelligence (AI) is increasingly popular, and educators are paying increasing attention to it. For students, learning AI helps them better cope with emerging societal, technological, and environmental challenges. This theory of planned behavior (TPB)-based study developed a survey questionnaire to measure behavioral intention to learn AI ( $n = 682$ ) among primary school students. The questionnaire was administered online, and it measured responses to five TPB factors. The five factors were (1) self-efficacy in learning AI, (2) AI readiness, (3) perceptions of the use of AI for social good, (4) AI literacy, and (5) behavioral intention. Exploratory factor analysis and a subsequent confirmatory factor analysis were used to validate this five-factor survey. Both analyses indicated satisfactory construct validity. A structural equation model (SEM) was constructed to elucidate the factors' influence on intention to learn AI. According to the SEM, all factors could predict intention to learn AI, whether directly or indirectly. This study provides new insights for researchers and instructors who are promoting AI education in schools.

**Keywords:** Artificial intelligence, Self-efficacy, Readiness, Social good, Literacy, Behavioral intention

### 1. Introduction

Artificial intelligence (AI) is rapidly developing, and it has become an integral part of our everyday lives. AI, which is ubiquitously adopted in many computing devices, is changing how we search for information, communicate with others, and make everyday arrangements (Lin et al., 2021). As an emerging field (regarded as part of the fourth industrial revolution), AI has also impacted the domain of education, potentially engendering a fourth education revolution (Roll & Wylie, 2016; Seldon & Abidoye, 2018). AI has been used in education for a range of purposes—including in administrative support systems (Sellar & Gulson, in press); intelligent tutoring systems (Ma, Adesope, Nesbit, & Liu, 2014; VanLehn, 2011); adaptive learning systems (Nakic, Granic, & Glavinic, 2015); and social assistive robots, used to make learning more engaging for preschool children (Fridin, 2014). The range of AI applications will only continue to grow, and the technology is likely to inspire new educational practices (So et al., 2020). However, as noted by Zawacki-Richter, Marín, Bond, and Gouverneur (2019) in their comprehensive review, studies on AI in education have focused only on higher education and on AI system development; research on how AI is taught and learned is therefore required.

Current K-12 students will face an AI-powered future, which is likely to demand greater creativity, critical thinking, and aptitude with technology (Aoun, 2017; Hwang & Fu, 2020); these can be enhanced through learning about AI. However, very little research has been conducted on learning AI among K-12 students (Zawacki-Richter et al., 2019). Most previous studies on AI in a K-12 context have been focused on the pedagogical use of AI systems—in, for example, intelligent tutoring systems—rather than the learning of AI itself. Although these systems have had demonstrable effectiveness (Ma et al., 2014), the role of the computer has evolved from that of being the tutor to that of a tutee or teachable agent (Chin, Dohmen, & Schwartz, 2013; Matsuda, Weng, & Wall, 2020). Regardless of the pedagogical model employed, students have used AI to learn something else; students have rarely learned about AI itself. Meanwhile, the need to develop AI curricula for younger students to acquire AI knowledge has begun to emerge (Knox, 2020). In China, for example, K-12 AI textbooks have been published, and curricular frameworks on AI have been formulated (Knox, 2020; Qin, Ma, & Guo, 2019; Tang & Chen, 2018); pilot testing of the AI curricula is ongoing. In general, the AI curricula in China covers the following: everyday applications of AI, how AI can solve problems (such as the early diagnosis of diseases), the core concepts underlying AI (e.g., data representation, machine learning, visual recognition, and the algorithm), and how to write code (Qin et al., 2019; Tang & Chen, 2018).

Against this background, this study investigated how primary school students learn AI, focusing specifically on their intention towards learning AI. This study adopted the theory of planned behavior (TPB) and developed a survey questionnaire. The AI curriculum had the dual aims of preparing students for an AI-powered workplace and to encourage students to consider AI as a possible future career (Qin et al., 2019). The first aim was advanced by promoting basic AI literacy. In general, greater AI literacy makes students more willing and able to engage with new technologies (Corbeil & Valdes-Corbeil, 2007; Parasuraman & Colby, 2015) and less fearful of an AI-powered world (Wang & Wang, 2019). The second aim was advanced by motivating students to continue learning AI. In general, such behavioral intention is partially formed by behavioral beliefs regarding first, the consequences of an action and second, one's ability to perform the action (Ajzen, 2012). Furthermore, students are more willing to learn AI if their ability to learn AI is fostered and if they understand the benefits of learning AI (Keller, 2010; Fishbein & Ajzen, 2010). This implies that designers of an AI curriculum must ensure (1) an appropriate level of difficulty and (2) ample illustrations with meaningful examples. In particular, meaningful illustrative examples can motivate students through illustrating AI's contributions to society. In general, such purposeful learning empowers students and encourages student participation (Yeager & Bundick, 2009). Computer skills curricula have had an aim of helping students understand computing's contributions to for social good (Goldweber et al., 2011). However, empirical studies between computing for social good and students' intention to learn computing seems lacking. Social good may be associated with behavioral intention to learn AI. Thus, the goal of this study is to include social good as purpose of AI learning to propose a framework guided by TPB to investigate the influential factors affecting behavioral intention towards learning AI. To achieve the goal, we develop and validate a contextualized survey.

## 2. Literature review

### 2.1. Behavioral intention to learn AI

The TPB (Ajzen, 1985) was developed from the theory of reasoned action (Fishbein & Azjen, 1975). Fishbein and Azjen (2010) conceptualized human behavior as reasoned action that follows from behavioral intention (BI); BI was, in turn, conceptualized as being based on "the information or beliefs people possess about the behavior under consideration" (p. 20). Such information is typically acquired through mass media or formal education. However, the same piece of information may be interpreted differently by different people depending on their individual-level traits, such as their personality and demographic characteristics. These differences mark those beliefs that determine BI. Fishbein and Azjen noted three types of individual-level beliefs: attitude towards behaviors (ATB), subjective norms (SN), and perceived behavioral control (PBC). These factors have consistently accounted for the many variances in learners' behavioral intentions in many empirical studies (Ajzen, 2012). Findings in the literature have jointly indicated that an individual is more likely to perform behaviors that are perceived to (1) yield positive outcomes, (2) be normatively desirable, and (3) involve controllable behavioral processes and outcomes. The TPB has been applied in numerous contexts, such as in technology adoption (Cheon, Lee, Crooks, & Song, 2012), health care (Chau & Hu, 2002), and e-commerce and business (Liao, Chen, & Yen, 2007). These studies have allowed social scientists to understand the influences on people's intention to use technology.

In the field of education, the TPB has been adopted by studies examining the intention to use technology, both in students and teachers (Cheon et al., 2012; Mei, Brown, & Teo, 2018). For example, Cheng, Chu, and Ma (2016) investigated students' attitude towards e-collaboration, and they noted that SN and PBC significantly predicted their intention to engage in e-collaboration. With regard to AI education, however, TPB-based research has yet to be conducted despite the increasing popularity of K-12 AI education. Thus, this study used the TPB to investigate the psychological factors that influence primary students' perceptions of and intentions towards learning AI. The findings help educators create favorable conditions for learning AI in the primary school classroom.

### 2.2. Background factors

As noted by Fishbein and Ajzen (2010), BI is shaped by background factors, which are categorized into individual (e.g., personality), societal (e.g., education, age and gender), and epistemic (e.g., knowledge and ways of thinking) factors. An individual is technologically literate if they know how technology works and how to use technology to solve problems (Moore, 2011)—including using technology to acquire, interpret, and apply knowledge (Davies, 2011). Thus, this study defines a student as AI literate if they know what constitutes AI and know how to apply AI to different problems. According to the TPB (Fishbein & Ajzen, 2010), AI literacy is

foundational to the behavioral, normative, and control beliefs that would consequently predict the BI (i.e., see Table 4, H1–H4). Previous studies have demonstrated that perceived technology literacy predicts (1) effort expectancy in e-learning (Mohammadyari & Singh, 2015) and (2) teachers' and students' intention to engage in mobile learning (Jong et al., 2018; Mac Callum, Jeffrey, & Kinshuk, 2014). Mei et al. (2018) also demonstrated that perceived technological literacy positively predicts learning intention, specifically in preservice teachers' intention to use technology in language education. In addition, Rubio, Romero-Zaliza, Mañoso, and de Madrid (2015) reported gender differences in the BI of university students towards coding after an introductory programming course. Their study also noted that this gender difference was eliminated by including the physical computing approach in course design. In general, their study elucidated the factors influencing BI and how TPB can be used to improve BI in students and teachers of technology-based courses.

### 2.3. Attitudes and behavioral intention

BI is influenced by an individual's ATB. ATB is defined as an individual's favorable or unfavorable feelings towards a psychological object (Fishbein & Ajzen, 2010). Ajzen (2012) stated that the favorableness associated with a behavior is largely based on the evaluative beliefs formed regarding the behavior's consequences. As mentioned, education aims to foster (1) citizens who contribute to social good and (2) members of the workforce who are well-equipped for the future workplace (Duncan & Sankey, 2019; Lo, 2010; White, 2010). Such aims have been incorporated into the present-day AI curriculum in China (Qin et al., 2019; Tang & Chen, 2018). In the context of computer science education, few empirical studies have investigated the influence on intention to learn AI from perceptions of AI's contributions to society; this is despite the fact that computing for social good is a recognized curriculum emphasis (Goldweber et al., 2011) and the strong ethical concern among AI scholars regarding the use of AI for social good (Bryson & Winfield, 2017; Floridi et al., 2018). Drawing from general educational research on occupational aspiration in adolescents, Yeager and Bundick (2009) observed a greater willingness to learn when their participants had goals that were directed towards something greater than themselves (Chang et al., 2020; Huang et al., 2019; Jong, Lee, & Shang, 2013; Lan et al., 2018). The idea of learning to use knowledge to serve others aligns with the philosophy of the Confucius which posits education as a means to perfect one's self in service of others (Basharat, Iqbal, & Bibi, 2011). By contrast, learning disengagement is likely if one learns merely to perform well in standardized tests (Dong et al., 2020; Jong et al., 2008; Taylor et al., 2013). Thus, the motivation to learn AI for social good can be considered as an attitude towards learning (Webb, Green, & Brashear, 2000). Webb et al. (2000) developed a scale for measuring an individual's attitude towards helping others, which is similar to our scale; such an attitude is strongly associated with the intention and willingness to act (Briggs, Peterson, & Gregory, 2010). According to the preceding analysis, this study posits that curriculum designed to promote learning for social good could strengthen students' intention to learn. Whether or not an AI curriculum that illustrates AI for social good will shape students' favorable ATB such that they could predict students' intention to learn AI remains to be investigated (i.e., H9 in Table 4).

The instrumental value of any curriculum is also to get students ready for the future. Among students with a favorable ATB, whether a given pedagogical technique prepares them well for the future requires empirical verification. The technology readiness index (TRI) was proposed by Parasuraman (2000) to measure one's propensity to use technology for a given set of goals. In the TRI, greater readiness indicates greater perceived control over a given piece of technology and a greater likelihood to use it often. The sense of readiness was deduced by the users from the knowledge and the confidence they possess (i.e., H3 and H6 in Table 4). Technology readiness has been used to explain the adoption of new technology (Parasuraman & Colby, 2015) and use of technology (Godoe & Johansen, 2012). Considering these findings in the literature, this study hypothesized that greater technological readiness predicts greater intention to learn AI. Specifically, the present study adapted items from Parasuraman's (2000) TRI to measure AI readiness in primary school students.

The sense of being ready should be considered as a self-oriented positive attitude, and one's positive attitude towards technology use can explain one's intention to use it (Chiu, 2017; Teo & Tan, 2012). Thus, using AI for social good may contribute to the participants' readiness to use it (i.e., H8 in Table 4). For example, Bertot, Jaeger, and Grimes (2010) illustrate how technology could promote transparency (a form of social good) that shape government adoption towards e-service. The TRI has been studied with the TPB in the context of e-commerce (Grandon, & Ramirez-Correa, 2018), where the TRI was employed as a background factor explaining how innovativeness changes the significance of PBC in predicting intention to adopt e-commerce. Furthermore, Chen and Li (2010) noted that intention to use e-services is predicted by ATB, SN, and PBC, which are, in turn, predicted by a combined factor of TRI. In both of the aforementioned studies, the TRI has been treated as a background personality trait that positively influences the intention to use technology. However, both studies had adult participants, for whom the TRI could be considered a background factor. However, among young learners



and especially for an emerging discipline like AI, technology readiness is more likely to be an outcome of learning, and such technological readiness, in turn, contributes to intention to learn (H10 in Table 4).

## 2.4. Perceived behavioral control and behavioral intention

PBC refers to one's perceived capability of performing a behavior (Ajzen, 1991). According to Ajzen (2002), PBC is conceptually similar to Bandura's concept of self-efficacy, defined as the perceived ease or difficulty in performing a behavior. The feeling of certainty and confidence in successfully executing a behavior under examination constitutes the core items that are frequently used to indicate measure self-efficacy (Fishbein & Ajzen, 2010). However, Ajzen (2002) also noted that beliefs regarding self-efficacy and beliefs regarding the controllability of a behavior can be two distinctive factors of PBC (see also Rhodes & Courneya, 2004). Nonetheless, Ajzen's (2002) analysis on this issue has pointed out that the self-efficacy factor is a stronger predictor for intention. Furthermore, studies commonly measure self-efficacy or confidence but not both (Zhang, Wei, Sun, & Tung, 2019).

Studies have demonstrated that in students, confidence in learning predicts continuous learning (Lee, 2010) and the intention to use technology as a learning tool (Garland & Noyes, 2005). In many TPB studies, self-efficacy is a commonly adopted PBC scale, and participants with self-efficacy have been noted to have greater BI (Rhodes & Courneya, 2004) (H7 in Table 4).

TPB-based studies have rarely investigated the relationship between PBC factors (e.g., self-efficacy) and ATB. Nonetheless, Yildiz's (2018) study shows that students' technology and communication self-efficacy contributes to their flipped learning readiness, which in turn predict their attitudes toward programming. It provides some support that in students, self-efficacy in learning AI predicts sense of readiness and predicts attitudes towards learning AI for social good and their sense of readiness (i.e., H5 & H6 in Table 4).

In sum, we hypothesized that for primary school students, the TPB could explain intention to learn AI. A structural model of students' BI and the variables that influence their intention to learn AI was developed (see Table 4 and Figure 1). Our research questions were as follows: (1) Is the 5-factor survey for primary school students' perception of AI learning valid and reliable? (2) Are the hypothesized relationships (H1-H 10) among the factors supported?

## 3. Method

### 3.1. Participants

Convenience sampling was used to enroll participants ( $N = 682$ , 52.05% male) in Beijing, China. The students were in the third to the sixth grades, with an average age of 9.87 years ( $SD = 0.97$  years). The school arranged for the participants to be enrolled in an AI course covering basic AI knowledge. Specifically, the course covered the history of AI, applications of AI (e.g., in image and voice recognition, content recommendation, and machine learning), and the ethical use of AI. As noted in classroom observations, the participants learned about basic AI concepts and data representations; they also participated in the hands-on use of AI products and discussions on the use of AI products. Students spent an average of 6.04 h ( $SD = 2.56$  h) on AI-related learning activities. The students were invited by their teachers to voluntarily respond to an online survey in the classroom at the end of the semester. The students took approximately 15 minutes to complete the survey. They were instructed to respond to each item by choosing the option that best described their level of agreement.

### 3.2. Instruments

This study's survey was based on five constructs, some of which were adapted from previous studies and others comprised self-constructed items. Answers were scored on a 4-point Likert scale from 1 (*strongly disagree*) to 4 (*strongly agree*). The first part of the survey collected background data (grade, gender, age, and hours spent on AI learning). The second part of the survey measured student confidence in AI, AI readiness, perceptions of using AI for social good, AI literacy, and BI to engage in AI learning. The finalized items are presented in Table 1. The following is a brief description of the five constructs of the survey.

**Self-efficacy in learning AI** was adapted from Song and Keller's (1999) confidence scale ( $\alpha = 0.70$ ), which was initially designed to measure students' confidence in the context of computer-mediated instruction. The items measured students' "self-efficacy varying in their degree of difficulty" (Fishbein & Ajzen, 2010, p.158). Specifically, the items measured students' confidence in their understanding, in how far they will succeed should they put in effort, and in their understanding of both advanced material and the basic concepts. In this study, greater confidence indicated greater self-efficacy (Fishbein & Ajzen, 2010) in meeting the learning objectives of the AI class.

**AI readiness** was developed from the optimism subscale of the TRI (Parasuraman, 2000). AI readiness is the student's perceived level of comfort with the use of AI technology in their everyday lives. Students with greater AI readiness favor the adoption of AI technology. The original scale had 10 items, among which six items were adapted for use in the present study (for the 10 items,  $\alpha = 0.78$ ).

**AI for social good** comprised five self-constructed items. It measured students' beliefs regarding the use of AI knowledge to solve problems and improve people's lives. The items indicated students' awareness of one purpose of learning AI.

**AI literacy** comprised five items. The items were developed based on the primary school's AI curriculum. AI literacy measured students' perception of their understanding of AI and of their general ability to use AI in their everyday lives.

**Behavioral intention** comprised adaptations of three of the four items in Park, Nam, and Cha (2012). Their study investigated university students' BI to be engaged in mobile learning (for the four items,  $\alpha = 0.91$ ). Furthermore, one more item was used in this scale. That item was adapted from Liaw, Huang, and Chen (2007), who investigated the BI to use e-learning.

The survey was reviewed by five professors in the fields of computer engineering and educational technology. The survey was then revised based on their comments. Subsequently, two teachers from the participating schools modified the wording of the survey's questions to ensure that students were able to understand the items.

### 3.3. Data analysis

This study's data analysis proceeded in three phases. In the first phase, the participants were randomly assigned to two subsamples. One subsample comprised approximately one-third of the participants ( $n = 217$ , 55.76% male); it was used for exploratory factor analysis (EFA). The other subsample comprised the remaining participants ( $n = 465$ , 50.32% male); it was used for confirmatory factor analysis (CFA) and SEM.

Prior to the analyses, univariate and multivariate normality tests were conducted for the entire data set. With respect to univariate normality, we noted that no measured item had a skewness (range:  $-0.989$  to  $-2.148$ ) and kurtosis (range:  $0.336$  to  $5.149$ ) that were greater than the requisite maximum values of  $|3|$  and  $|8|$ , respectively (Kline, 2011). With respect to multivariate normality, Mardia's coefficient is the standard indicator. This value should be less than  $(k[k+2])$ , where  $k$  is the number of observed variables (Raykov & Marcoulides, 2008). For this study, the coefficient value was  $521.392$ , which was less than the requisite maximum of  $22 \times 24 = 528$ . Multivariate normality was thus satisfied.

EFA was conducted using SPSS (version 25) to clarify the structure of the subscales. Principal axis factoring analysis and the direct oblimin rotation method were applied to extract the factors. Items with cross loadings or factor loadings of  $< 0.5$  were omitted. Alpha reliabilities were computed for all factors and items. Pearson correlation analysis was used to analyze the relationship between the factors. Subsequently, CFA was conducted to verify the construct validity of the instrument. Structural equation modeling (SEM) was then used for hypothesis testing in Amos for Structural Equation Modeling (version 23).

## 4. Results

### 4.1. Exploratory factor analysis of the measurement model

Table 1 summarizes the EFA results, including the mean, standardized deviation, factor loadings, and alpha reliabilities. The EFA extracted 22 items with factor loadings greater than 0.5 in the final version of the 5-factor

measurement model. The Kaiser–Meyer–Olkin value was 0.910, and the value for Bartlett’s test of sphericity was 4008.041 ( $df = 231$ ,  $p < .001$ ). These results indicated that the five factors had good explanatory power with respect to perception of AI learning.

The five factors explained 69.97% of the variance in perception of AI learning; they were self-efficacy in learning AI (four items,  $\alpha = 0.88$ ), AI readiness (five items,  $\alpha = 0.88$ ), perceptions of the use of AI for social good (five items,  $\alpha = 0.92$ ), AI literacy (four items,  $\alpha = 0.91$ ), and BI (four items,  $\alpha = 0.90$ ). The overall  $\alpha$  value was 0.95, which suggested that these factors had satisfactory reliability and they were suitable for measuring perceptions of AI learning.

*Table 1.* Exploratory factor analysis results for intention to learn AI ( $n = 217$ )

Item	Factor loading
Self-efficacy, $\alpha = 0.88$ , $M = 3.53$ , $SD = 0.61$	
C3 I am certain I can understand the most difficult material presented in the AI class.	0.77
C1 I feel confident that I will do well in the AI class.	0.76
C4 I am confident I can learn the basic concepts taught in the AI class.	0.73
C2 I believe that I can succeed if I try hard enough in the AI class.	0.58
Readiness, $\alpha = 0.88$ , $M = 3.62$ , $SD = 0.49$	
RE2 It is much more convenient to use the products and services that use the latest AI technologies.	0.81
RE6 I feel confident that AI technologies will follow the instructions I give.	0.71
RE1 AI technology gives people more control over their daily lives.	0.69
RE3 I prefer to use the most advanced AI technology available.	0.66
RE4 I like AI technology that allows me to tailor things to fit my own needs.	0.66
Social good, $\alpha = 0.92$ , $M = 3.63$ , $SD = 0.54$	
SG1 I wish to use my AI knowledge to serve others.	0.84
SG2 AI can be used to help disadvantaged people.	0.83
SG4 AI combined with design thinking can enhance my ability to help others.	0.70
SG3 AI can promote human well-being.	0.67
SG5 The use of AI should aim to achieve common good.	0.62
Literacy, $\alpha = 0.91$ , $M = 3.59$ , $SD = 0.58$	
L2 I can use AI-assisted voice recognition software to search for information.	0.84
L1 I know that AI can be used to recognize images.	0.78
L4 I am able to use online AI translation tools.	0.74
L3 I can interact with AI assistants via speech recognition (e.g., Siri, DuerOS).	0.68
Behavioral intention, $\alpha = 0.90$ , $M = 3.51$ , $SD = 0.67$	
BI3 I will continue to acquire AI-related information.	0.94
BI2 I will keep myself updated with the latest AI applications.	0.89
BI4 I intend to use AI to assist with my learning.	0.62
BI1 I will continue to learn AI.	0.52

#### 4.2. Correlations among the factors

Pearson correlation coefficients were calculated to investigate the relationships among the five factors. As noted in Table 2, these factors were significantly and positively correlated (from  $r = 0.54$  to  $r = 0.63$ ).

*Table 2.* Correlations in the measured model ( $n = 217$ )

	1	2	3	4	5
1. Self-efficacy	(0.71)	0.62***	0.54***	0.58***	0.60***
2. Readiness		(0.71)	0.59***	0.49***	0.58***
3. Social good			(0.73)	0.63***	0.63***
4. Literacy				(0.76)	0.55***
5. Behavioral intention					(0.76)

*Note.* \*\*\* $p < .001$ . Items on the diagonal are the square roots of the average variance extracted; off-diagonal elements are the correlation estimates.

### 4.3. Confirmatory factor analysis of the measurement model

The CFA further confirmed the construct validity and the structure of the measurement model. As detailed in Table 3, all item parameters were statistically significant.

The model had good fit:  $\chi^2/df = 2.89$  ( $< 5.0$ ), RMSEA = 0.064 ( $< 0.08$ ), SRMR = 0.044 ( $< 0.05$ ), GFI = 0.90 ( $> 0.90$ ), TLI = 0.94 ( $> 0.90$ ), and CFI = 0.95 ( $> 0.90$ ) (Hair, Black, Babin, Anderson, & Tatham, 2010). More generally, these results indicated that the survey items had good construct validity.

Moreover, the examination of the composite reliability (CR) of each sub-scale was greater than 0.70, and the average variance extracted (AVE) met or exceeded the value of 0.50: Self-efficacy in learning AI (CR = 0.80, AVE = 0.50), AI readiness (CR = 0.83, AVE = 0.50), AI for social good (CR = 0.85, AVE = 0.54), AI literacy (CR = 0.85, AVE = 0.58), and behavioral intention (CR = 0.84, AVE = 0.58), indicating satisfactory reliability and convergent validity of each sub-scale (see Hair et al., 2010). Discriminant indexes (See Table 2) were computed based on the AVEs.

Table 3. Confirmatory factor analysis results for intention to learn AI ( $n = 465$ )

Scale	Item	Mean	SD	Unstandardized estimate	Standardized estimate	<i>t</i> -value
Self-efficacy	C1	3.35	0.79	1	0.81	-
	C2	3.51	0.67	0.85	0.82	19.85***
	C3	3.23	0.85	1.01	0.76	18.01***
	C4	3.47	0.73	1.00	0.88	21.81***
Readiness	RE1	3.60	0.64	0.99	0.84	19.28***
	RE2	3.62	0.61	0.96	0.86	19.81***
	RE3	3.53	0.70	1	0.77	-
	RE4	3.47	0.74	1.09	0.80	18.13***
Social good	RE6	3.36	0.81	0.97	0.64	14.17***
	SG1	3.54	0.70	1	0.89	-
	SG2	3.56	0.68	0.92	0.83	23.99***
	SG3	3.58	0.66	0.87	0.82	23.60***
	SG4	3.52	0.70	0.95	0.84	24.42***
Literacy	SG5	3.58	0.69	0.88	0.80	22.24***
	L1	3.58	0.66	1	0.86	-
	L2	3.60	0.62	0.96	0.87	22.75***
	L3	3.55	0.68	0.93	0.77	19.19***
Behavioral intention	L4	3.64	0.65	0.88	0.76	18.86***
	BI1	3.44	0.76	0.94	0.86	25.94***
	BI2	3.42	0.76	1	0.91	-
	BI3	3.40	0.76	0.95	0.86	26.25***
	BI4	3.39	0.81	0.71	0.60	14.65***

Note. \*\*\* $p < 0.001$ .

### 4.4. SEM for hypotheses testing

SEM was used for hypothesis testing. The SEM model had good fit:  $\chi^2/df = 2.91$  ( $< 5.0$ ), RMSEA = 0.064 ( $< 0.08$ ), SRMR = 0.044 ( $< 0.05$ ), GFI = 0.90 ( $> 0.90$ ), TLI = 0.94 ( $> 0.90$ ), CFI = 0.95 ( $> 0.90$ ) (Hair et al., 2010). As shown in Table 4, eight out of ten hypotheses were confirmed, indicated that AI literacy significantly predicts Self-efficacy in learning AI and social good. Self-efficacy is a significant predictor for social good, AI readiness, and behavioral intention. Social good significantly predicts AI readiness and behavioral intention. AI readiness significantly predicts behavioral intention. The estimated standardized path coefficients are presented in Figure 1. The findings show that most hypothesized relationships among the sub-scales were supported.

Table 4. Hypotheses testing results from SEM

Hypothesis	Path	Unstandardized estimate	Standardized estimate	<i>t</i> -value	Hypotheses supported?
H1	Literacy $\rightarrow$ Self-efficacy	0.68	0.59	11.67***	Yes
H2	Literacy $\rightarrow$ Social good	0.43	0.39	7.71***	Yes
H3	Literacy $\rightarrow$ Readiness	0.05	0.05	1.12	No
H4	Literacy $\rightarrow$ Behavioral intention	-0.08	-0.06	-1.34	No

H5	Self-efficacy—> Social good	0.42	0.43	8.30***	Yes
H6	Self-efficacy —> Readiness	0.30	0.36	6.85***	Yes
H7	Self-efficacy —> Behavioral intention	0.40	0.37	6.52***	Yes
H8	Social good —> Readiness	0.45	0.52	9.23***	Yes
H9	Social good —> Behavioral intention	0.37	0.34	5.20***	Yes
H10	Readiness —> Behavioral intention	0.32	0.25	3.55***	Yes

Note. \*\*\* $p < 0.001$ .

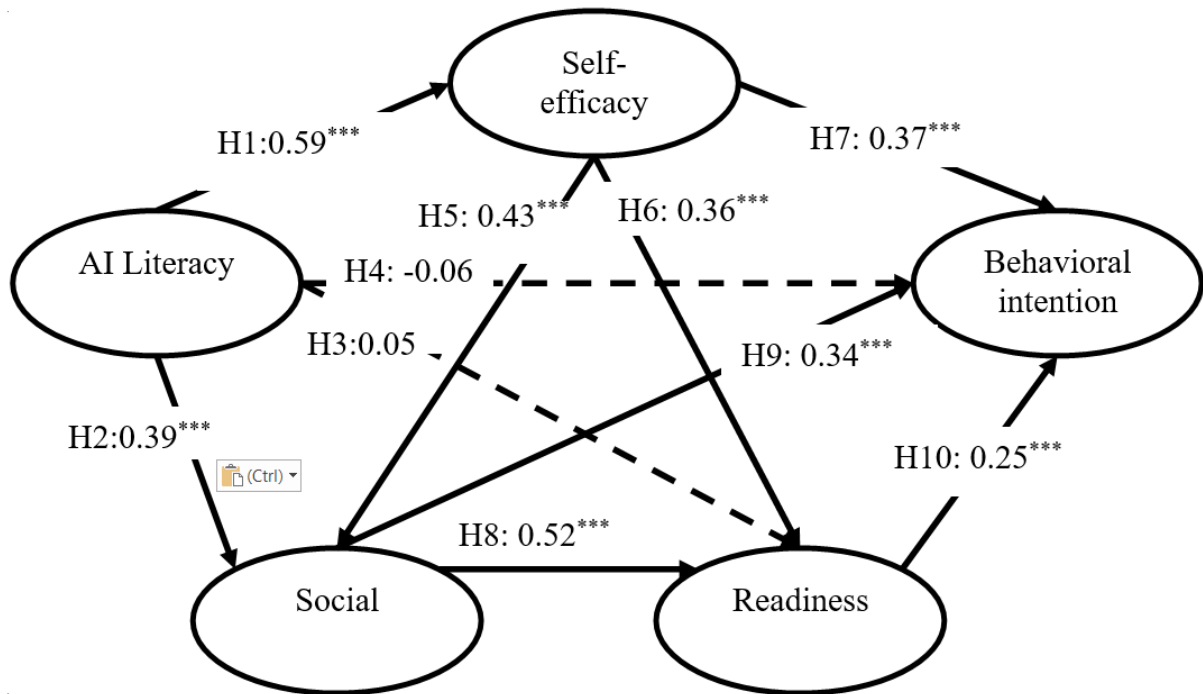


Figure 1. Structural model of the measured factors. Note. \*\*\* $p < 0.001$

## 5. Discussion and conclusions

Although AI has received much recent attention in the higher-education context, it has rarely been explored in the K-12 context (Zawacki-Richter et al., 2019). Considering the need to prepare young students for an AI-powered workplace, students' intention to learn AI must be investigated. This TPB-based study surveyed primary school students' intention to learn AI (Fishbein & Ajzen, 2010). Fishbein and Ajzen recognized the context dependence in people's assignment of weights to the factors pertaining to attitude, norms, and perceived control. Considering the context of the new AI curriculum, this study chose students' perceived AI literacy as a background factor; perceived use of AI for social good and readiness for the AI-powered world as their ATB; and confidence in learning AI as their PBC. These factors were hypothesized to predict students' intention to learn AI. The findings from 682 primary school students (Grades 3 to 6) in Beijing indicated that intention to learn AI was influenced by self-efficacy in learning AI, AI readiness, and perceived use of AI for social good. The background factor, AI literacy, only influenced students' self-efficacy in AI and perceptions of the use of AI for social good directly. The findings of this study are generally congruent with those of studies using the TPB, as formulated by Fishbein and Ajzen. Similar to the many previous TPB-based studies (Mei et al., 2018; Mohammadyari & Singh, 2015; Zhang et al., 2019), this study noted the TPB to be a useful theoretical framework for identifying factors that contribute to BI. Our findings suggest that to foster strong BI towards learning AI, developers of AI curricula should pay attention to students' ATB and PBC. The implications of this study are discussed as follows.

First, this study employed EFA and CFA to establish a valid and reliable five-factor survey that measures students' perception of learning AI. This survey can be used in future research on how curriculum design influences BI. Within the Web of Science Core Collection database, we identified 5470 studies containing the search term "theory of planned behavior." However, further separate searches within this result using "elementary OR primary" did not return with any study. This study therefore could further enrich the

applicability of TPB in primary school contexts. For primary school students, their first experience of learning AI in formal educational ought to prepare them for the future AI-powered workplace. This study's survey indicated that the curriculum used by the participants could provide positive experiences and foster students' readiness and BI for learning AI in the future. Specifically, the mean scores for all measured factors were >3.5; a score of 3 constituted the neutral point in the 4-point scale that these factors were scored on. Our findings also elucidated the factors that influence BI.

The background factor of AI literacy was defined as the knowledge of AI that students acquired from the curriculum. In accordance with Fishbein and Ajzen's (2010) model that depicts knowledge or information as predicting ATB and PBC, and current studies applied to technology-based teaching and learning (Mei et al., 2018; Mohammadyari & Singh, 2015; Mac Callum et al., 2014); AI literacy is a significant predictor of the students' self-efficacy and the social good. However, it did not predict students' readiness and BI directly. This indicates that AI literacy is not a sufficient condition for being ready to learn or use AI. Gaining PBC (i.e., self-efficacy) and a belief that AI contributes to social good are necessary. Designers of AI curricula must, therefore, pay special attention to these aspects.

Self-efficacy was the most important factor that directly predicted students' BI, AI readiness, and perceptions about the use of AI for social good. While self-efficacy can predict students' BI in learning as indicated by past research (Lee, 2010; Garland & Noyes, 2005), the finding also points to the importance of PBC, and it contributes to the further understanding of the relationships between PBC and ATB in the context of learning AI. Little research has explicated the PBC may predict ATB and the implication of this study could be informative for future research involving curriculum design for an emerging field of study. In such a context, addressing students' self-efficacy could be crucial in shaping their evaluative beliefs about learning the subject matter.

Furthermore, perceptions of learning AI for social good significantly predicted students' readiness to learn AI and intention to learn AI. This suggests that AI curricula should allow students to solve real-world problems to illustrate how AI can be used to benefit others; this encourages students to delve deeper into AI. Such an emphasis on the use of AI for social good is congruent with current trends in computer science education (Goldweber et al., 2011; Bryson & Winfield, 2017; Floridi et al., 2018). Students are likely to regard the promotion of social good as a positive outcome of learning AI, which, in turn, fosters BI (Fishbein & Ajzen, 2010) and purpose in learning (Yeager & Bundick, 2009). Pedagogically, teachers should use examples and hands-on applications to illustrate how AI can be used for social good, thereby stimulating students' intention to learn AI. These strategies are reflected in the present-day AI curriculum in China (Qin et al., 2019).

Student readiness also predicts intention to learn. Greater readiness, as measured in this study, reflected a more positive perception of how useful AI is. Greater readiness was interpreted as another positive consequence of learning AI. AI literacy predicted readiness, not directly but only indirectly, through self-efficacy and the perception of using AI for social good; this finding is congruent with a previous finding that readiness grants an individual a sense of control by helping them use technology flexibly and efficiently (Parasuraman & Colby, 2015). This implies that readiness is not immediately obvious to students. However, if students are confident that they can learn and use AI for social good, they feel more ready to use AI. Furthermore, students who perceive AI to be useful have a greater intention to learn it; this finding is congruent with those of earlier studies (Mei et al., 2018; Yildiz, 2018).

In conclusion, the emergence of AI has greatly changed society and technology, and education must reform itself accordingly (Aoun, 2017; Seldon & Abidoye, 2018). Students ought to be prepared to learn AI early in their education. According to Fishbein and Ajzen (2010), people tend to deliberate on their actions when encountering a novel situation, of which the emergence of AI is one; the result of such deliberation, in turn, forms the cognitive foundation for future decisions. We recommend for educators to foster self-efficacy and emphasize the potential use of AI for social good. In doing so, students are more likely to have greater intention to learn AI, and they can thus be better prepared for an AI-powered future.

## 6. Limitations

This study has several limitations. First, this study was limited to the primary school students in Beijing, China. Further research should examine and compare K-12 students' AI learning from other cities or countries and levels of students. Second, this study considered only positive attitudes in its measures of intention to learn AI. However, adverse psychological factors, such as anxiety towards AI (Wang & Wang, 2019), should also be considered—especially considering the increasing adoption of AI education in primary schools. Third, the TPB

postulates three conceptually independent determinants of intentions (i.e., ATB, SN, and PBC). These factors have accounted for a large proportion of the variances in the variables of many previous studies. This study, however, did not include SN as a variable. Future research should investigate SN as a potential facilitating condition (Mei et al., 2018). Fourth, this study measured PBC using only self-efficacy towards AI. Rhodes and Courneya's (2004) study discovered that adding a phrase "If I wanted to" to such items could influence the effects of ATB, SN and PBC on BI. It is suggested that this phrase should be included in the items in the future. Future studies can also consider including control items that can affect PBC.

## Acknowledgment

The survey items for measuring technological readiness in this study were adapted from the Technology Readiness Index, which is copyrighted by A. Parasuraman and Rockbridge Associates, Inc. 2000. This scale may be duplicated only with written permission from the original authors.

## References

- Ajzen, I. (1985). From intentions to actions: A Theory of planned behavior. In J. Kuhl & J. Beckmann (Eds.), *Action control* (pp. 11–39). Berlin, Germany: Springer.
- Ajzen, I. (1991). The Theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179–211.
- Ajzen, I. (2002). Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. *Journal of Applied Social Psychology*, 32(4), 665–683.
- Ajzen, I. (2012). The Theory of planned behavior. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 438–459). London, UK: Sage.
- Aoun, J. E. (2017). *Robot-proof: Higher education in the age of artificial intelligence*. Cambridge, MA: MIT Press.
- Basharat, T., Iqbal, H. M., & Bibi, F. (2011). The Confucius philosophy and Islamic teachings of lifelong learning: Implications for professional development of teachers. *Bulletin of Education and Research*, 33(1), 31–46.
- Bertot, J. C., Jaeger, P. T., & Grimes, J. M. (2010). Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. *Government Information Quarterly*, 27(3), 264–271.
- Briggs, E., Peterson, M., & Gregory, G. (2010). Toward a better understanding of volunteering for nonprofit organizations: Explaining volunteers' pro-social attitudes. *Journal of Macromarketing*, 30(1), 61–76.
- Bryson, J., & Winfield, A. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. *Computer*, 50(5), 116–119.
- Chang, S. C., Hsu, T. C., Kuo, W. C., & Jong, M. S. Y. (2020). Effects of applying a VR-based two-tier test strategy to promote elementary students' learning performance in a Geology class. *British Journal of Educational Technology*, 51(1), 148–165.
- Chau, P. Y., & Hu, P. J. H. (2002). Investigating healthcare professionals' decisions to accept telemedicine technology: An Empirical test of competing theories. *Information & Management*, 39(4), 297–311.
- Chen, S. C., & Li, S. H. (2010). Consumer adoption of e-service: Integrating technology readiness with the theory of planned behavior. *African Journal of Business Management*, 4(16), 3556–3563.
- Cheng, E. W. L., Chu, S. K. W., & Ma, C. S. M. (2016). Tertiary students' intention to e-collaborate for group projects: Exploring the missing link from an extended theory of planned behavior model. *British Journal of Educational Technology*, 47(5), 958–969.
- Cheon, J., Lee, S., Crooks, S., & Song, J. (2012). An Investigation of mobile learning readiness in higher education based on the theory of planned behavior. *Computers & Education*, 59(3), 1054–1064.
- Chin, D. B., Dohmen, I. M., & Schwartz, D. L. (2013). Young children can learn scientific reasoning with teachable agents. *IEEE Transactions on Learning Technologies*, 6(3), 248–257.
- Chiu, T. K. F. (2017). Introducing electronic textbooks as daily-use technology in schools: A Top-down adoption process. *British Journal of Educational Technology*, 48(2), 524–537.
- Corbeil, J. R., & Valdes-Corbeil, M. E. (2007). Are you ready for mobile learning? *Educause Quarterly*, 30(2), 51–58.
- Davies, R. S. (2011). Understanding technology literacy: A Framework for evaluating educational technology integration. *TechTrends*, 55(5), 45–52.

- Dong, A. M., Jong, M. S. Y., & King, R. (2020). How does prior knowledge influence learning engagement? The mediating roles of cognitive load and help-seeking. *Frontiers in Psychology*, 11, 591203.
- Duncan, C., & Sankey, D. (2019). Two conflicting visions of education and their consilience. *Educational Philosophy and Theory*, 51(14), 1454–1464. doi:10.1080/00131857.2018.1557044
- Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An Introduction to theory and research*. Reading, MA: Addison-Wesley.
- Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The Reasoned action approach*. London, UK: Psychology Press.
- Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds & Machines* 28, 689–707. doi:10.1007/s11023-018-9482-5
- Fridin, M. (2014). Storytelling by a kindergarten social assistive robot: A Tool for constructive learning in preschool education. *Computers & Education*, 70, 53–64.
- Garland, K., & Noyes, J. (2005). Attitudes and confidence towards computers and books as learning tools: A Cross-sectional study of student cohorts. *British Journal of Educational Technology*, 36(1), 85–91.
- Godoe, P., & Johansen, T. (2012). Understanding adoption of new technologies: Technology readiness and technology acceptance as an integrated concept. *Journal of European Psychology Students*, 3(1), 38–52.
- Goldweber, M., Davoli, R., Little, J. C., Riedesel, C., Walker, H., Cross, G., & Von Konsky, B. R. (2011). Enhancing the social issues components in our computing curriculum: computing for the social good. *ACM Inroads*, 2(1), 64–82.
- Grandon, E. E. & Ramirez-Correa, P. (2018). Managers/Owners' innovativeness and electronic commerce acceptance in Chilean SMEs: A Multi-group analysis based on a structural equation model. *Journal of Theoretical and Applied Electronic Commerce Research*, 13(3), 1–16.
- Hair Jr, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2010). SEM: An Introduction. *Multivariate data analysis: A global perspective* (pp. 629–686). Upper Saddle River, NJ: Pearson Education.
- Huang, C. Q., Han, Z. M., Li, M. X., Jong, M. S. Y., Tsai, C. C. (2019). Investigating students' interaction patterns and dynamic learning sentiments in online discussions. *Computers & Education*, 140, Article 103589.
- Hwang, G. J., & Fu, Q. K. (2020). Advancement and research trends of smart learning environments in the mobile era. *International Journal of Mobile Learning and Organisation*, 14(1), 114–129.
- Jong, M. S. Y., Chan, T., Hue, M. T., & Tam, V. (2018). Gamifying and mobilizing social enquiry-based learning in authentic outdoor environments. *Educational Technology & Society*, 21(4), 277–292.
- Jong, M. S. Y., Lee, J. H. M., & Shang, J. J. (2013). Educational use of computer game: Where we are and what's next? In R. Huang, Kinshuk, & J. M. Spector (Eds.), *Reshaping Learning: Frontiers of Learning Technology in a Global Context* (pp. 299–320). Heidelberg: Springer.
- Jong, M. S. Y., Shang, J. J., Lee, F. L., & Lee, J. H. M. (2008). Harnessing games in education. *Journal of Distance Education Technologies*, 6(1), 1–9.
- Keller, J. M. (2010). *Motivational design for learning and performance: The ARCS model approach*. Boston, MA: Springer.
- Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.
- Knox, J. (2020). Artificial intelligence and education in China. *Learning, Media and Technology*. doi:10.1080/17439884.2020.1754236
- Lan, Y. J., Botha, A., Shang, J. J., & Jong, M. S. Y. (2018). Technology enhanced contextual game-based language learning. *Educational Technology & Society*, 21(3), 86–89.
- Lee, M. C. (2010). Explaining and predicting users' continuance intention toward e-learning: An Extension of the expectation–confirmation model. *Computers & Education*, 54(2), 506–516.
- Liao, C., Chen, J. L., & Yen, D. C. (2007). Theory of planning behavior (TPB) and customer satisfaction in the continued use of e-service: An Integrated model. *Computers in Human Behavior*, 23(6), 2804–2822.
- Liaw, S. S., Huang, H. M., & Chen, G. D. (2007). Surveying instructor and learner attitudes toward e-learning. *Computers & Education*, 49(4), 1066–1080.
- Lin, P. Y., Chai, C. S., Jong, M. S. Y., Dai, Y., Guo, Y., & Qin, J. (2021). Modeling the structural relationship among primary students' motivation to learn artificial intelligence. *Computers & Education: Artificial Intelligence*, 2, 100006.
- Lo, J. T. Y. (2010). The Primary social education curricula in Hong Kong and Singapore: A Comparative study. *Research in Comparative and International Education*, 5(2), 144–155.



- 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.
- Mac Callum, K., Jeffrey, L., & Kinshuk. (2014). Comparing the role of ICT literacy and anxiety in the adoption of mobile learning. *Computers in Human Behavior*, 39, 8–19.
- Matsuda, N., Weng, W., & Wall, N. (2020). The Effect of metacognitive scaffolding for learning by teaching a teachable agent. *International Journal of Artificial Intelligence in Education*, 30, 1–37. doi:10.1007/s40593-019-00190-2
- Mei, B., Brown, G. T. L., & Teo, T. (2018). Toward an understanding of preservice English as a foreign language teachers' acceptance of computer-assisted language learning 2.0 in the People's Republic of China. *Journal of Educational Computing Research*, 56(1), 74–104.
- Mohammadyari, S., & Singh, H. (2015). Understanding the effect of e-learning on individual performance: The Role of digital literacy. *Computers & Education*, 82, 11–25.
- Moore, D. R. (2011). Technology literacy: The Extension of cognition. *International Journal of Technology and Design Education*, 21(2), 185–193.
- Nakic, J., Granic, A., & Glavinic, V. (2015). Anatomy of student models in adaptive learning systems: A Systematic literature review of individual differences from 2001 to 2013. *Journal of Educational Computing Research*, 51(4), 459–489.
- Parasuraman, A. (2000). Technology readiness index (TRI): A Multiple-item scale to measure readiness to embrace new technologies. *Journal of Service Research*, 2, 307–320.
- Parasuraman, A., & Colby, C. L. (2015). An Updated and streamlined technology readiness index: TRI 2.0. *Journal of Service Research*, 18(1), 59–74.
- Park, S. Y., Nam, M. W., & Cha, S. B. (2012). University students' behavioral intention to use mobile learning: Evaluating the technology acceptance model. *British Journal of Educational Technology*, 43(4), 592–605.
- Qin, J. J., Ma, F. G., & Guo, Y. M. (2019). *Foundations of artificial intelligence for primary school*. Beijing, CN: Popular Science Press.
- Raykov, T., & Marcoulides, G. A. (2008). *An Introduction to applied multivariate analysis*. New York, NY: Taylor & Francis.
- Rhodes, R. E., & Courneya, K. S. (2004). Differentiating motivation and control in the theory of planned behavior. *Psychology, Health & Medicine*, 9(2), 205–215.
- Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education*, 26(2), 582–599.
- Rubio, M. A., Romero-Zaliz, R., Mañoso, C., & de Madrid, A. P. (2015). Closing the gender gap in an introductory programming course. *Computers & Education*, 82, 409–420.
- Seldon, A., & Abidoye, O. (2018). *The Fourth education revolution*. London, UK: Legend Press Ltd.
- Sellar, S., & Gulson, K. N. (in press). Becoming information centric: The Emergence of new cognitive infrastructures in education policy. *Journal of Education Policy*. doi:10.1080/02680939.2019.1678766
- So, H. J., Jong, M. S. Y., & Liu, C. C. (2020). Computational thinking education in the Asian Pacific region. *The Asia-Pacific Education Researcher*, 29(1), 1–8.
- Song, S. H., & Keller, J. M. (1999). The ARCS Model for the design of motivationally adaptive computer-mediated instruction. *Journal of Educational Technology*, 14(1), 119–134.
- Tang, X., & Chen, Y. (2018). *Fundamentals of artificial intelligence*. Shanghai, CN: East China Normal University.
- Taylor, E. L., Taylor, P. C., & Chow, M. (2013). Diverse, disengaged and reactive: A Teacher's adaptation of ethical dilemma story pedagogy as a strategy to re-engage learners in education for sustainability. In N. Mansour & R. Wegerif (Eds.), *Science Education for Diversity* (pp. 97–117). Dordrecht, Netherlands: Springer.
- Teo, T., & Tan, L. (2012). The theory of planned behavior (TPB) and pre-service teachers' technology acceptance: A Validation study using structural equation modeling. *Journal of Technology and Teacher Education*, 20(1), 89–104.
- VanLehn, K. (2011). The Relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221.
- Wang, Y. Y., & Wang, Y. S. (2019). Development and validation of an artificial intelligence anxiety scale: an initial application in predicting motivated learning behavior. *Interactive Learning Environments*, doi:10.1080/10494820.2019.1674887
- Webb, D. J., Green, C. L., & Brashear, T. G. (2000). Development and validation of scales to measure attitudes influencing monetary donations to charitable organizations. *Journal of the Academy of Marketing Science*, 28(2), 299–309.

White, J. (2010). *The Aims of education restated*. London, UK: Routledge.

Yeager, D. S., & Bundick, M. J. (2009). The Role of purposeful work goals in promoting meaning in life and in schoolwork during adolescence. *Journal of Adolescent Research*, 24(4), 423–452.

Yildiz, H. D. (2018). Flipped learning readiness in teaching programming in middle schools: Modelling its relation to various variables. *Journal of Computer Assisted Learning*, 34(6), 939–959.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 39. doi:10.1186/s41239-019-0171-0

Zhang, F., Wei, L., Sun, H., & Tung, L. C. (2019). How entrepreneurial learning impacts one's intention towards entrepreneurship: A Planned behavior approach. *Chinese Management Studies*, 13(1), 146–170.

# Gender Differences in Cognitive Load when Applying Game-Based Learning with Intelligent Robots

Beyin Chen<sup>1</sup>, Gwo-Haur Hwang<sup>2\*</sup> and Shen-Hua Wang<sup>3</sup>

<sup>1</sup>Department of Information Technology, Ling Tung University, Taiwan // <sup>2</sup>Bachelor Program in Industrial Technology, National Yunlin University of Science and Technology, Taiwan // <sup>3</sup>Department of Information Management, Ling Tung University, Taiwan // byc@teamail.ltu.edu.tw // ghhwang0424@gmail.com // happy0487587@gmail.com

\*Corresponding author

**ABSTRACT:** The application of artificial intelligence (AI) in education is now widespread, and the use of robots in education has demonstrated a positive influence on students' behavior and development. However, the use of emerging technologies usually results in cognitive load, especially for elementary school students whose learning capacity has not yet been established. In addition, students of different genders have different physical, psychological and learning characteristics, so gender differences affect cognitive load. Cognitive load can be divided into two types: positive cognitive load and negative cognitive load. Usually, positive cognitive load results in good learning performance while negative cognitive load results in bad learning performance. Therefore, we use the cognitive load theory to define learning efficiency as the co-impact of learning performance and cognitive load. We take game-based intelligent robots for Chinese idiom learning as an example, and explore the impacts of gender differences on elementary school students. To achieve these aims, this study combined games and Zenbo robots, and applied them to educate elementary school students in the use of Chinese idioms. Secondly, this study conducted an experiment and analyzed the experimental results. The participants were 24 fourth-grade elementary school students from the central region of Taiwan. Results showed that this system is more beneficial for boys as their cognitive load was significantly lower. Boys' learning performance was also better, although the difference did not reach significance. Furthermore, learning efficiency for boys was significantly higher. Reasons for these results are explained.

**Keywords:** Artificial intelligence, Cognitive load theory, Game-based learning, Gender differences, Robots

## 1. Introduction

The work of artificial intelligence (AI) is dedicated to solving cognitive problems that are usually related to human intelligence. There is no denying that the education sector has been significantly affected by AI. AI in education (AIEd) is now widely used by learners and educators (Chen, Xie, & Hwang, 2020; Hwang, Xie, Wah, & Gašević, 2020). For example, Hwang, Sung, Chang, and Huang (2020) developed an adaptive learning system and explored its associated mathematical anxiety and cognitive load. Their experimental results showed that the proposed approach helped the low achievers successfully complete the learning tasks. A form of AI in education is the use of social robots (Papadopoulos, Lazzarino, Miah, & Weaver, 2020). Papadopoulos et al. (2020) pointed out that the use of robots in education has demonstrated a positive influence on the behavior and development of students, especially in the areas of problem-solving skills (Barak & Zadok, 2009) and teamwork practice (Varney, Janoudi, Aslam, & Graham, 2012), increased motivation to learn (Kubilinskiene, Zilinskiene, Dagiene, & Sinkevicius, 2017), and enhancement of participation (Rusk, Resnick, Berg, & Pezalla-Granlund, 2008).

Microsoft founder and chairman Bill Gates predicted that the robot industry will become the next hot area (Gates, 2007). Due to the rapid development of the future robot industry, the urgency of promoting robot education or related cross-disciplinary education is expected to increase day by day (Hsiao & Huang, 2012). Westlund et al. (2017) pointed out that robots leverage human means of communicating, such as speech, movement and nonverbal cues, including gaze, gestures, and facial expressions, in order to interface with us in more natural ways. Their study also pointed out that the emotional expressiveness of the robot's speech might modulate children's learning. Therefore, leveraging robots to help elementary school students learn idioms is a feasible solution.

On the other hand, the theoretical basis of the influence of games on learning is derived from children's cognitive development psychology. Cognitive development theory states that the growth of learners in different cognitive stages needs to be completed by the maturity and transformation of the previous stage (Piaget, 1964). Cai, Yan, Yang, and Wang (2012) also pointed out that if game situations can be used to help children to learn joyfully, the children's focus quality during the learning process can last longer, and then the stage's maturity and transformation can be reached.

In the past, several studies also integrated games and robots. For example, the robot Mindstorms integrates system simulation and program manipulation into games. It creates a broad field of vision for learning to help users learn in more depth, and has become a classic example of joyful learning (Resnick & Ocko, 1991; Rusk et al., 2008). Liu and Lin (2009) also pointed out that the impacts of combining games and smart robots will be different from those of traditional simple game-based learning, which creates a broader research area for joyful learning.

Emerging technologies and game-based learning can help students increase their learning interest (Ng'ambi, 2013). However, they usually result in cognitive load, especially for younger students whose learning capacity has not yet been established. For emerging technologies, Zhong, Zheng, and Zhan (2020) examined the effects of virtual and physical robots (VPR) used in different learning stages (simple session/complex session) in a robotics programming course. They found that significant difference existed in engineering design ability and cognitive load, no matter whether in simple or complex learning sessions. Huang, Shadiev, and Hwang (2016) explored the effectiveness of applying speech-to-text recognition (STR) technology during lectures in English on the cognitive load of non-native English speaking students. The result showed that lectures in English caused less cognitive load for low ability EFL students when they used STR-texts. For game-based learning, Liao, Chen, and Shih (2019) investigated how the use of an instructional video and collaboration influenced the intrinsic motivation and cognitive load of students learning Newtonian mechanics within a digital game-based learning (DGBL) environment. While collaborative DGBL promoted intrinsic motivation, the results for cognitive load showed that the use of an instructional video in collaborative DGBL significantly reduced both intrinsic and extraneous cognitive loads. Javora, Hannemann, Stárková, Volná, and Brom (2019) examined the effects of a holistic and appealing visual design of a learning game. Based on cognitive-affective theory of learning with media and cognitive load theory, they found that visual design's influence on learning outcomes is mediated by (at least) two hidden variables: cognitive engagement and cognitive load.

However, Bevilacqua (2017) pointed out that gender differences in cognitive load have resulted in some differences in aspects of associated working memory systems that are relevant to cognitive load theory. He also indicated that if males and females process information differently in working memory, cognitive load levels will be different for males and females experiencing similar stimuli under certain conditions. Many studies have pointed out that in different learning environments, gender may have an impact on cognitive load. For example, Christophel and Schnotz (2017) explored the correlations between cognitive load and competences. They found that the correlations differed for female and male participants. Wong, Castro-Alonso, Ayres, and Paas (2015) explored gender effects when learning manipulative tasks from instructional animations and static presentations. They found that the cognitive load of females and males have reverse results.

However, in our survey of the literature, we found that past studies exploring gender differences in cognitive load seldom investigated robots. Therefore, our study explored the impacts of gender differences in cognitive load on applying game-based learning by intelligent robots. Sweller (1998) defined cognitive load as the amount of load generated when a specific task is applied to an individual's cognitive system. Its management is embodied within the framework of the cognitive load theory (CLT) (Sweller, 2005). CLT has been proven to be a theory of great value for instructional design (Paas, van Gog, & Sweller, 2010). If the measurement of cognitive load is combined into the related aspects of CLT, it is likely to be more accurate and complete. The related aspects include cognitive load, learning performance and learning efficiency (Paas & van Merriënboer, 1993). Among these three aspects, learning efficiency reflects the co-impact of learning performance and cognitive load. The goal of our study is that when the educational robots are used by elementary school students, learning efficiency can be better exerted in response to the influence of gender differences.

Based on the above, the research questions of this study are described as follows:

RQ1: Does gender have a significant impact on the cognitive load of elementary school students in learning idioms?

RQ2: Does gender have a significant impact on the learning performance of elementary school students in learning idioms?

RQ3: Does gender have a significant impact on the learning efficiency of elementary school students in learning idioms?

## **2. Literature review**

### **2.1. Artificial intelligence (AI) in education**

The rapid development of computing and information processing technology has accelerated the development and application of AI, which aims to enable computers to perform tasks via simulating human intelligent behaviors, such as reasoning, analysis, and decision-making (Hwang, Xie, Wah, & Gašević, 2020). From the perspective of precision education, it emphasizes the need to provide prevention and intervention measures for learners by analyzing their learning conditions or behaviors (Hart, 2016). Chen, Xie, and Hwang (2020) indicated that AIED is widely used by learners and educators nowadays, and involves various tools and applications, for example, intelligent tutoring systems, teaching robots, and adaptive learning systems. AIED supports learning in traditional classes and workplaces by combining AI with various learning sciences such as education, psychology, linguistics and neuroscience, and aims to stimulate and promote AI-driven educational application (Luckin, Holmes, Griffiths & Forcier, 2016).

With the development of AI technology, the “natural language processing (NLP)” feature of robots is playing a pivotal role. NLP is the ability of computers and cloud-based applications (apps) to communicate with humans in their own natural languages, such as Chinese or English (Smith, Haworth, & Žitnik, 2020). It refers to building computational tools that analyze and represent human language at human communication complexity levels (Liddy, 2001). Several innovative educational products have claimed their adoption of AI-enabled techniques to facilitate learning performance, with applications ranging from chatbots with NLP techniques (Crossley, Allen, Snow, & McNamara, 2016). In our research, we used robots’ NLP feature to interact with elementary school students to learn Chinese idioms.

### **2.2. Application of robots in elementary school education**

Emerging technologies are transforming society and inspiring technological innovation in previously un-thought-of practices, beliefs and perceptions (Ng’ambi, 2013). Robots are one of the increasingly popular emerging technologies that have the potential to influence students’ learning (Kucuk & Sisman, 2017). Several studies have reported that the use of robots helps children engage more in their learning activities. For example, Wu, Wang, and Chen (2015) used the robot’s facial expressions, gestures, and movements to generate various forms of communication and interaction with the students, thereby helping the elementary school students learn English. The experimental results showed that the students’ learning experience, motivation and achievement improved. Hsiao et al. (2019) used robot-based practices to develop an activity that incorporated the 6E (engage, explore, explain, engineer, enrich, and evaluate) model to improve elementary school students’ learning effects. With the 6E model, the instructor facilitated the students’ hand-made process by strengthening the connection between life experience, learning content, learning characterization, and interdisciplinary knowledge.

However, in the past, some studies indicated that gender may affect the use of educational robots for elementary school students. For example, Master, Cheryan, Moscatelli, and Meltzoff (2017) indicated that young girls had less interest and self-efficacy in technology compared with boys in elementary school. Therefore, considering the issue of gender differences in cognitive load is necessary.

### **2.3. Cognitive load theory (CLT) and the impacts of gender on cognitive load**

Cognitive load (CL) in e-learning has been explored for many years and its management is embodied within the framework of the CLT (Sweller, 2005). The CLT emphasizes that all novel information is initially processed by working memory which has capacity and duration limitations; the information is then stored in long-term memory which is unlimited. The aim of instructional design should be to reduce unnecessary working memory loads, and free the capacity for learning-related processing to accommodate the limited capacity of working memory (Mutlu-Bayraktar, Cosgun, & Altan, 2019). CL has two meanings. One is a causal dimension that reflects the interaction between personal traits and task traits. The other is a dimension of measurement that describes the measurable aspects of mental load (ML), mental effort (ME), and performance (PE) (Krell, 2017; Paas & van Merriënboer, 1994). ML is task-related and refers to the cognitive ability required to handle task complexity. Conversely, ME is subject-related and reflects the cognitive ability invested when a person is working on a task. The relationship between PE and CL is unknown. For example, a subject may have the same number of correct answers (i.e., PE) in one test, but may require a different amount of ME (Paas, Tuovinen, Tabbers, & van Gerven, 2003). Paas and van Merriënboer (1993) proposed a method that provides a tool to

correlate ME measurements of PE. This approach defines high learning efficiency as low ME to achieve high PE, while low learning efficiency is defined as high ME to achieve low PE.

However, gender differences may result in differences in cognitive load (Bevilacqua, 2017). In the past, many researchers explored gender difference in cognitive load. For example, Wong, Castro-Alonso, Ayres, and Paas (2015) explored gender effects when learning manipulative tasks from instructional animations and static presentations. The results showed that for females, on such tasks, using animations might have clearer advantages in managing their cognitive load rather than static presentations. For males, the reverse strategy may be more effective. Hwang, Hong, Cheng, Peng, and Wu (2013b) conducted experiments to explore the gender differences in cognitive load for sixth-grade students. The results showed that girls had a higher cognitive load and more competition anxiety from synchronous types of competitive games.

Although many studies have explored the influence of gender on cognitive load in various learning environments in the past, we have rarely found research on the influence of gender in cognitive load in the learning environment of educational robots. Therefore, we believe that it is a worthwhile topic to explore the impact of gender on the cognitive load of elementary school students in the learning environment of educational robots. These aspects include cognitive load (CL), learning performance (PE) and learning efficiency (LE). We measured cognitive load (CL) by defining the questions of the CL questionnaire as the sum of the questions of the mental load (ML) questionnaire and the mental effort (ME) questionnaire (Hwang, Yang, & Wang, 2013a). It was expected that the application of CLT could make the issue of the impact of gender difference on cognitive load better understood from a wider range of perspectives.

### 3. System development

#### 3.1. System architecture

In this study, we combined intelligent robots and game-based learning to enhance students' learning. Among various types of robots, we chose Zenbo, which was launched by ASUS (i.e., ASUSTeK Computer Inc. in Taiwan) in January 2017. It is 62 cm tall and weighs 10 kg. It can display 24 expressions such as happiness, shyness and vitality, as well as body movements such as raising its head, walking and rotation (see <https://www.youtube.com/watch?v=zW23nLYRWck>). The shape, sound and body movements are all designed to simulate the cute image of a two-year-old boy. Zenbo is a type of robot with the NLP feature, which is one of the important technologies of modern AI. These characteristics make Zenbo highly suitable for elementary school students (Chen, Hwang, Wang, & Peng, 2018). Based on these, we developed the "Zenbo Idiom Learning System (ZILS)" to help elementary school students learn idioms. This system is divided into two modules. They are the "idiom learning" module and the "reviewing with games" module. A unit includes seven idioms and takes 40 minutes to complete. The system architecture is shown in Figure 1.

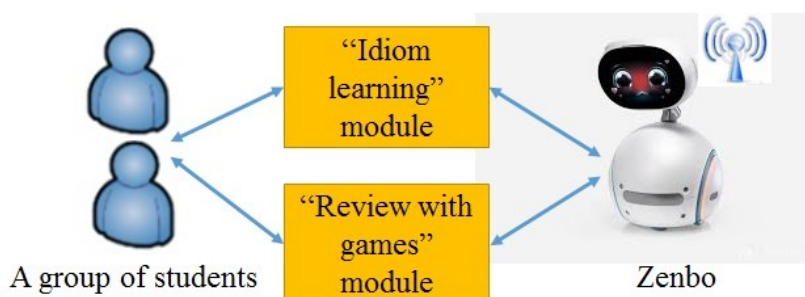


Figure 1. System architecture

#### 3.2. The "idiom learning" module

Since idioms are defined as linguistic expressions, their overall meaning cannot be predicted from the meanings of the constituent parts (literal meaning) (Kovecses & Szabco, 1996). Therefore, idiom interpretation is important for elementary school students. In addition, the Chinese poet Yu (1994) stated that idioms are a unique feature of the Chinese language, often with historical stories and philosophical significance. Yu (1994) pointed out that, at present, many people who write Chinese do not use idioms. He also stated that the decline in the use of idioms shows that classical Chinese is being forgotten and its cultural significance is shrinking. To arouse

students' interest in using idioms, besides understanding their meaning, the classical allusions of idioms and idiom sentence-making are important.

Based on these reasons, the “idiom learning” module mainly teaches idiom interpretation, the classical allusion of idioms, and idiom sentence-making. In the “idiom learning” module, the content of the seven idioms is displayed and played by Zenbo’s screen and voice. The process is interspersed with pictures and texts. We chose the idioms according to themes such as numbers, animals, colors, people, etc. Each unit deals with a different theme, for example: for the first unit, we selected animals as the theme and then we selected seven idioms. Each idiom comprises four Chinese words. We translated the seven idioms into English as follows: “to see a bow reflected in a cup as a snake,” “the mantis stalks the cicada, unaware of the oriole behind,” “the tortoise and the hare,” “to mend the pen after the sheep are lost,” “like a fish back in water,” “if you ride a tiger, it’s hard to get off” and “to refrain from shooting at the rat for fear of breaking the vases.” Table 1 provides the classical allusion of the idiom of “to refrain from shooting at the rat for fear of breaking the vases.”

*Table 1. An example of the classical allusions*

The idiom	“to refrain from shooting at the rat for fear of breaking the vases”
The classical allusion	There is a story in the book of Han which tells of a rich man who was a lover of antiques and who had a large collection. Among them, there was a rare vase made of jade. Due to the vase having exquisite workmanship and historical value, he loved it dearly. One night, a mouse jumped into the jade vase and wanted to eat some leftovers inside. The rich man happened to see this scene. He was very annoyed. In a rage, he took a stone and smashed it on the mouse. Of course, the mouse was killed, but the precious jade vase was also broken. The loss of the vase pained the man greatly and he deeply regretted his own thoughtlessness, which brought him this unrecoverable loss. He now realized that anyone who cares for the present while overlooking consequences is apt to bring disaster upon himself. So, he warned people by saying, “Do not burn your house to get rid of a mouse.”

In the “idiom learning” module, we designed its homepage to show seven animals representing the seven idioms as shown in Figure 2. According to the students’ selection, the system enters the interpretation, the classical allusion, and the sentence-making of the idiom. Zenbo will read aloud the text of this content. This design enables a smooth combination of pictures and speech, and matches the multimedia principles proposed by Mayer (2009). In addition, Mayer (2009) pointed out that the combination of words and pictures can improve the learning outcomes for students with low prior knowledge, so this design is also helpful for low level prior knowledge children. Figure 3 shows the screen of the classical allusion of the idiom “to mend the pen after the sheep are lost.”



*Figure 2. The homepage of the “idiom learning” module*



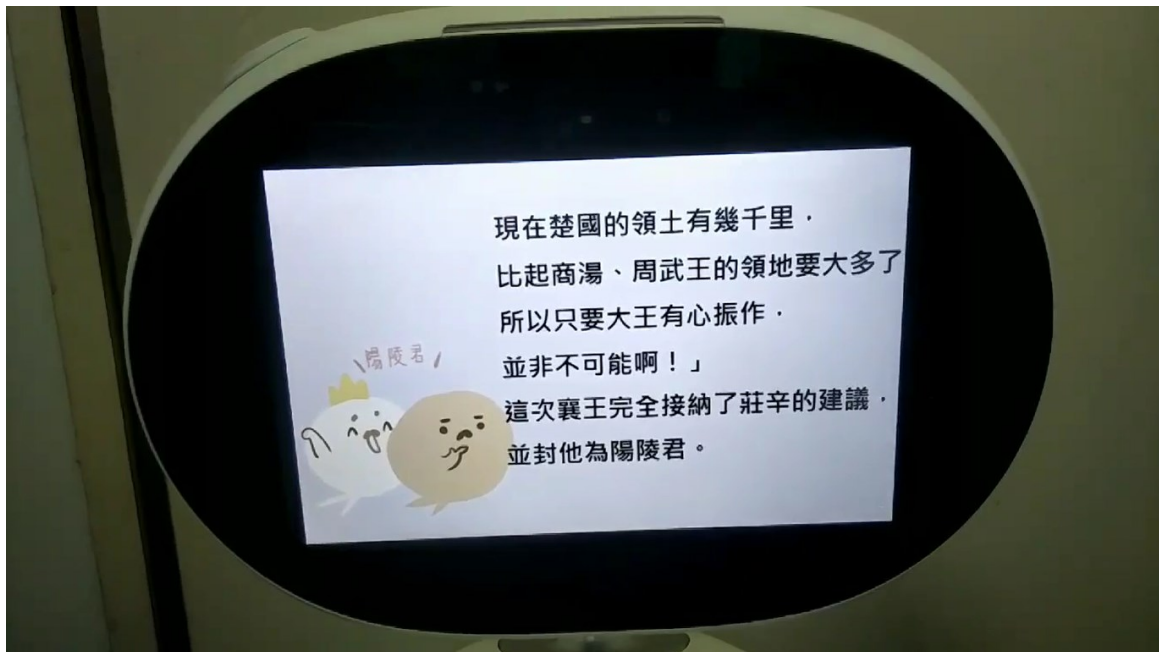


Figure 3. The classical allusion of the idiom “to mend the pen after the sheep are lost”

### 3.3. The “reviewing with games” module

The “reviewing with games” module reviews idioms using a game-based style. The module allows students to review the seven idioms by guessing them based on related pictures which correspond to the “words” of the idiom. Taking “to mend the pen after the sheep are lost” as an example, the first prompting picture may be a fence. The children will guess the idiom and say it out loud. If the answer is right, Zenbo will reply with a sentence such as, “Congratulations! Your answer is right.” If the answer is wrong, Zenbo will reply with a sentence such as, “Your answer is wrong. Try again!” In this case, a further prompting picture will appear. This picture may be a sheep. If the child’s answer is wrong again, Zenbo will reply with a sentence such as, “Your answer is wrong. Try for the final time!” Subsequently, the final prompting picture will appear. This picture may be a net. If the child still fails to answer the idiom completely, that idiom will be skipped. When all seven idioms have been reviewed, the failed ones will be repeated until the time is up. This “reviewing with games” module can deepen the impressions of the learned idioms.

To implement the “reviewing with games” module, we designed three prompting pictures for each idiom. Taking “to mend the pen after the sheep are lost” as an example, the three prompting pictures are a picture of a net, a picture of a sheep and a picture of a fence, as shown in Figure 4. Taking the picture of a net or a fence as an example, this is not a direct corresponding relationship with the word, but is an indirect correspondence, which can test students’ concentration ability. The embedded game elements combined with robots’ NLP function in the module “reviewing with games” facilitated students’ learning or attracted their learning attention.

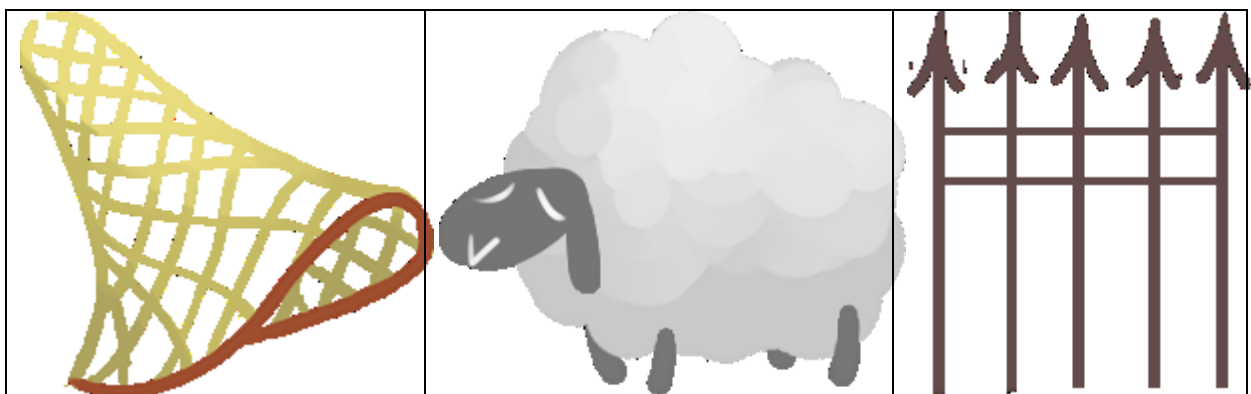


Figure 4. The three prompting pictures of the idiom “to mend the pen after the sheep are lost”



## 4. Research method

### 4.1. Research tools

In this study, the research tools include ZILS, the pre- and post-test questions, the questionnaire of cognitive load, and the SPSS statistical software. The 20 multiple-choice pre- and post-test questions were the same but in a different order. They were developed by the authors by extending the seven idioms in different aspects and expanding them into 20 questions which were compiled according to idioms appropriate for fourth-grade elementary school students in Taiwan. These 20 questions were reviewed and revised for reliability by two senior elementary school teachers, ensuring that they had expert validity. The cognitive load questionnaire was developed with reference to Hwang et al. (2013a). The questions were divided into two categories: mental load and mental effort. The Cronbach's  $\alpha$  of the questionnaires are .92 for mental effort, .81 for mental load and .89 for cognitive load, all of which surpass the suggested threshold value of .7. These questionnaires therefore have high reliability. A 5-point Likert scale was adopted. After the experiment was finished, SPSS19 was used to analyze the data. We have selected two questions from the 20 multiple-choice questions to demonstrate the reliability of assessing students' learning performance. The questions of the cognitive load questionnaire are also listed in Table 2.

Table 2. The two selected questions and the questions of the cognitive load questionnaire

Two selected questions	<ol style="list-style-type: none"> <li>1. Who is the related historical figure of the idiom "like a fish in water"? (A) Xiang Yu (B) Wen Tianxiang (C) Qu Yuan (D) Zhuge Liang</li> <li>2. Which of the following statements is best used to illustrate "social laziness"? (A) The mantis catches the cicada and the oriole is behind (B) Three monks have no water to drink (C) Three people must have my teacher (D) Three days of fishing and two days of drying the net</li> </ol>
Mental effort questions	<ol style="list-style-type: none"> <li>1. The learning process of "ZILS" caused me a lot of pressure.</li> <li>2. I had to put a lot of effort into completing the learning task of "ZILS".</li> <li>3. The learning process of "ZILS" was difficult to understand.</li> </ol>
Mental load questions	<ol style="list-style-type: none"> <li>1. The learning content of "ZILS" was difficult for me.</li> <li>2. It took me a lot of effort to answer the "ZILS" game questions.</li> <li>3. Answering "ZILS" game questions was very tiring.</li> <li>4. Answering "ZILS" game questions made me feel very frustrated.</li> <li>5. I didn't have enough time to answer "ZILS" game questions.</li> </ol>

### 4.2. Research architecture

In this study, the independent variable is gender. All the students took advantage of one lesson to learn the seven idioms using ZILS, so the control variables are teaching time and teaching materials. The explored questions are the impacts of gender differences on the elementary school students' related aspects of CLT, including cognitive load, learning performance and learning efficiency. Therefore, the dependent variables are cognitive load, learning performance and learning efficiency. The definition of learning efficiency referred to Paas and van Merriënboer (1993). The research architecture is shown in Figure 5.

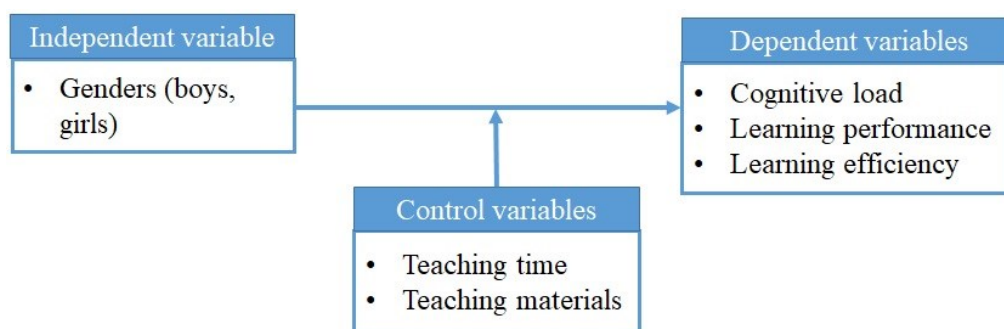


Figure 5. Research architecture

### 4.3. Experimental subjects

In this study, the experimental subjects are one class of fourth-grade students from an elementary school in the central region of Taiwan. There were 24 students in total including 12 boys and 12 girls. Figures 6 and 7 show scenes of the children using ZILS and filling out the questionnaires.



*Figure 6. Children using ZILS*



*Figure 7. Children filling out questionnaires*

#### 4.4. Research flow

The research flow is shown in Figure 8. First, we introduced the system to the students and the pre-test was conducted. This stage took 25 minutes. Second, the students used ZILS to learn the seven idioms. This stage took 40 minutes which included the use of 25 minutes for the “idiom learning” module and 15 minutes for the “reviewing with games” module. During the operation of the “reviewing with games” module, the children played in groups. Each student had at least one chance to interact with Zenbo. We adopted the style of grabbing the chance to answer. When one child had answered correctly, the chance was given to another who had not yet correctly answered. If no child grabbed the chance, we assigned someone who had not passed successfully. If all the children in a group had answered successfully, they were given the chance to answer repeatedly. This learning activity process involved interaction with Zenbo’s NLP features, and included game pictures and game rules which could drive the learning atmosphere and enhance the students’ learning effectiveness. Finally, the post-test and the cognitive load questionnaire were conducted. This stage took 25 minutes.

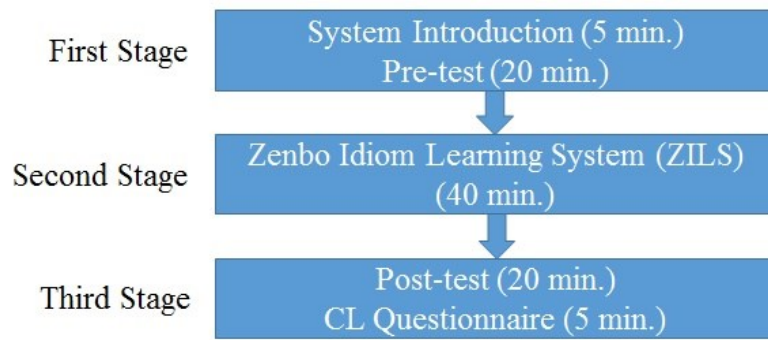


Figure 8. Research flow

### 5. Results

In this section, we present the results of the gender differences in the aspects of cognitive load, learning performance and learning efficiency. In the final part of this section, we explore the reasons to explain the results.

#### 5.1. Impact on cognitive load

To explore whether there were significant differences in the three indicators of cognitive load, mental effort and mental load, we conducted the independent samples *t* test. The results are shown in Table 3 where cognitive load is the weighted average of mental load and mental effort.

Table 3. The independent samples *t* test for analyzing mental effort, mental load, and cognitive load

Load categories	Sex	Number	Average	<i>SD</i>	<i>t</i>	Cohen's <i>d</i> (Effect size)
Mental Effort	Boys	12	1.425	0.458	2.859*	0.825
	Girls	12	2.283	0.934		
Mental Load	Boys	12	1.508	0.656	2.335*	0.674
	Girls	12	2.225	0.837		
Cognitive Load	Boys	12	1.464	0.506	2.844**	0.872
	Girls	12	2.238	0.795		

Note. \* $p < .05$ ; \*\* $p < .01$ .

From Table 3, it can be seen that boys rate significantly lower than girls for all aspects of mental effort ( $t = 2.859$ ,  $p < .05$ ,  $d = 0.825$ ), mental load ( $t = 2.335$ ,  $p < .05$ ,  $d = 0.674$ ) and cognitive load ( $t = 2.844$ ,  $p < .05$ ,  $d = 0.872$ ). Cohen (1988) has provided benchmarks to define small ( $d = 0.2$ ), medium ( $d = 0.5$ ), and large ( $d = 0.8$ ) effects. Accordingly, the effect sizes of the aforementioned results are all above medium, which suggested that they are acceptable. The average scores for boys are between 1.42 and 1.50 and those for girls are between 2.22 and 2.28. The gap of the average is close to 1. This phenomenon matches the results of Hwang et al. (2013b) who found that the cognitive load and competition anxiety of females are higher than those of males.

## 5.2. Impact on learning performance

To explore whether the learning performances of the different genders are significantly different, we conducted ANCOVA analysis. In the analysis, the pre-test was regarded as a co-variate, the post-test as a dependent variable, and gender as an independent variable. First, the interaction effects of the independent variable and the co-variate were observed. The results showed that the between-group and the pre-test of the ANCOVA analysis is not significant ( $F = 3.422, p = .078 > .05$ ). This means that the independent variable and the co-variate do not interactively affect the result, so the analysis could be continued. For the two genders, the ANCOVA analysis of excluding the impact of the pre-test is shown in Table 4.

Table 4. ANCOVA results of the two genders after excluding the impact of the pre-test

Groups	Number	Average	SD	Adjusted average	<i>F</i>	<i>p</i>
Boys	12	9.08	2.353	8.997	2.579	0.123
Girls	12	7.42	2.968	7.503		

From Table 4, it is seen that the adjusted average scores of the post-tests are 8.997 and 7.503 for the boys and girls respectively. It is seen that after excluding the impact of the pre-test, the two genders did not reach significant difference ( $F = 2.579, p = .123 > .05$ ). Although the difference is not significant, boys' learning performance is higher than that of girls.

## 5.3. The co-impact of normalized gain score and mental effort score

Paas and van Merriënboer (1993) proposed learning efficiency, which is measured by the value of the normalized performance score subtracting the mental effort score. In this study, we used the gain score to represent the performance score. The gain score is the score of the post-test subtracting the score of the pre-test. To normalize the gain score, we used the linear equation of  $A * X + B = Y$ , where  $X$  represents the gain score and  $Y$  represents the normalized gain score, as the normalized formula. Then we mapped the maximum normalized gain score to 5 (i.e., maximum score of the ME questionnaire) and the minimum normalized gain score to 1 (i.e., minimum score of the ME questionnaire). Sequentially, two equations were obtained. We solved the two equations and found  $A$  and  $B$ . In our case, for the 24 valid samples, the maximum value of the gain scores is 6 and the minimum value is (-3). The two equations are as follows:

$$\begin{aligned} 6 * A + B &= 5 \\ (-3) * A + B &= 1 \end{aligned}$$

By solving the two equations, the results are that  $A$  equals (4/9) and  $B$  equals (7/3). Finally, we normalized the gain score with the formula of  $(4/9) * X + (7/3) = Y$ . According to this formula and the learning efficiency equation of  $LE = Y - ME$ , we could calculate the normalized gain score for each child and then the learning efficiency could be calculated. Sequentially, the independent samples  $t$  test of the learning efficiency of the boys and girls could be analyzed. The results are shown in Table 5.

Table 5. The independent samples  $t$  test for analyzing learning efficiency

Sex	Number	Average	SD	<i>t</i>	Cohen's <i>d</i> (Effect size)
Boys	12	2.067	1.032	2.915**	0.842
Girls	12	0.575	1.441		

Note. \*\* $p < .01$ .

From the table, it can be seen that the boys' learning efficiency was significantly higher than that of girls ( $t = 2.915, p < .01, d = 0.842$ ). Accordingly, the effect size of the aforementioned results is large, which suggested that they are acceptable. This means that the instructional design benefits boys more than girls.

## 6. Discussion

From the above experimental results, it can be seen that the cognitive load of boys is significantly lower than that of girls. Although the difference in learning performance is not significant, that of boys is higher than that of girls. After further observation, the learning efficiency obtained after subtracting the mental effort score from the normalized gain score, the average score of boys is significantly higher than that of girls. These results represent that the use of educational robots to learn idioms for elementary school students is more beneficial for boys than



for girls. By observing the data, there were four students whose gain scores were negative. These four students were all girls. We therefore infer that for some girls, ZILS may cause disturbances to their learning. As the gain scores are all positive for boys, we infer that boys are generally more interested in emerging technologies and games. This result is similar to the finding of Hwang et al. (2013b) who found that for elementary school students, girls have more cognitive load and competition anxiety than boys.

The reasons behind these phenomena can be explained by referring to the literature. For example, Pauls, Petermann, and Lepach (2013) found that male students are more interested in visual and spatial aspects, while female students are more interested in the hearing aspect. Herlitz and Rehnman (2008) obtained similar results. The feelings of the boys when using ZILS may be related to the cute shape of the robot, the pictures in the system, and the placement of the robot in the classroom, which may bring changes to the classroom space. These are attractive to boys. Emerging technologies and games may cause girls learning interference. Webley (1981) also pointed out that boys explore new environments more frequently than girls. The author's argument also explains our results: boys may be more interested in exploring changes in the environment. Besides, Master et al. (2017) indicated that young girls had less interest and self-efficacy in technology compared with boys in elementary school. Ring (1991) also indicated that male students had greater self-confidence than female students in their ability to use courseware (and hence computers) as an effective learning tool. Dweck, Davidson, Nelson, and Enna (1978) indicated that females showed greater evidence of non-adaptive attributions and therefore were more adversely affected by failure. Therefore, the integration of emerging technologies and games into teaching will be of more benefit to boys.

With more in-depth discussion, our system has two major characteristics of game pictures and robotics technology. It brings changes in the look of the classroom space, with a pleasing appearance, which can arouse the interest of boys with strong sensory abilities such as vision and space. Girls with strong hearing ability may prefer to study quietly and have a greater cognitive load on general scientific and technological teaching activities. However, if science and technology teaching activities can be designed to have a strong spiritual level and have the connotation of deep learning in the process, it is likely to help girls reduce their cognitive load when using technology for learning. Therefore, although this research developed one unit only, in the future, two more units can be designed without duplication of idioms, and the difficulty of guessing the idioms in the three units can increase in order. For example, the first unit uses "character" as the reminder object to test children's memory of words. The second unit could use "idiom story" as the reminder theme to test students' memory and understanding of the storyline, while the third unit could use "idiom sentence-making" as a reminder theme to test students' understanding of the application of sentence-making. The game could be designed to interact with the robot in a complex manner, for example, when the idiom story is a reminder, not only could the picture be used, but the robot's NLP function could also be designed to interact with the students in the storyline picture. In the question/answer interaction, students would not only be guessing idioms, but also people and things in the story. Such learning content would not only improve the quality of the game elements, but would also make full use of the NLP function of the modern AI of robots. It may be able to help girls improve their learning effectiveness and reduce their cognitive load, thereby enhancing their learning efficiency.

Based on such future improvements, our system will be able to provide more real-time interaction and contextual learning to increase elementary school students' interest in learning Chinese idioms when compared with the past studies which have explored how to increase elementary school students' interest in learning Chinese idioms (Ku, Huang, & Hus, 2015; Wong, Chin, & Tan, 2010). The impacts of gender differences on other indicators of cognitive load, such as germane cognitive load (Lange & Costley, 2019) can be further verified.

## 7. Conclusions and future works

This study explored the topic of "gender differences in cognitive load on applying game-based learning by intelligent robots". To achieve this aim, this study developed a game-based learning system based on Zenbo robots, which uses the sound and the cute image of the Zenbo robot to attract elementary school students' interest in the learning content. During the learning process, the first module in ZILS is "idiom learning." It uses a homepage showing animal images, and lets Zenbo read out the screen's text to impress the students with the idiom story. The second module is "reviewing with games." It uses creative pictures from the game as a reminder for students to guess the idioms. This design can refresh the students' memory. We also explored the impacts of gender differences on the related aspects of CLT for elementary students in using the system. The results show that this system is more beneficial for boys. Their cognitive load is significantly lower. Their learning performance is also better, although it does not reach significance. Furthermore, the learning efficiency for them is significantly higher. From the literature, we also found that males and females have different learning

interests. Males are mainly interested in the visual and spatial aspects while females are mainly interested in the hearing aspect. These may be the reasons for the differences in the results.

In the future, if the system can be improved by increasing the game elements and the interaction with the robot's NLP function, it will be able to benefit girls and may have deeper findings. In addition, when using emerging technologies for E-learning, boys and girls have different physical, psychological and learning characteristics (Liao, Zhen, & Hwang, 2018). If more information of the learning process can be recorded and the impact on students' behavior can be explored, it should be possible to sort out more helpful rules and can serve as optimization criteria for the development of learning systems. Therefore, gender differences in behavior analysis in the environments of educational robots will be an important research issue that is worth exploring in the future, in order to optimize the system and improve students' learning performance.

## Acknowledgement

This study is supported by the Ministry of Science and Technology of Taiwan under contract number MOST 107-2511-H-275-001 and MOST 108-2511-H-224-006-MY3.

## References

- Barak, M., & Zadok, Y. (2009). Robotics projects and learning concepts in science, technology and problem solving. *International Journal of Technology and Design Education*, 19(3), 289–307.
- Bevilacqua, A. (2017). Commentary: Should gender differences be included in the evolutionary upgrade to cognitive load theory? *Educational Psychology Review*, 29(1), 189–194.
- Cai, J. S., Yan, R. C., Yang, J. J., & Wang, R. Y. (2012, May-June). *A Study of individual difference on learners' acceptance toward game-based instructional design of 3D animation*. Paper presented at the 16th Global Chinese Conference on Computers in Education, Pingtung, Taiwan.
- Chen, B., Hwang, G. H., Wang, S. H., & Peng, J. H. (2018, March). *The Development of an intelligent robot navigation system - A case study of Du-Xing elementary school located in north district of Taichung of Taiwan*. Paper presented at the 3rd International Conference on Digital Learning Strategies and Applications (DLSA), Hokkaido, Japan.
- Chen, X., Xie, H., & Hwang, G. J. (2020). A Multi-perspective study on artificial intelligence in education: Grants, conferences, journals, software tools, institutions, and researchers. *Computers & Education: Artificial Intelligence*, 1, 100005. doi:10.1016/j.caeai.2020.100005
- Christophel, E., & Schnotz, W. (2017). Gender-specific covariations between competencies, interest and effort during science learning in virtual environments. *Frontiers in Psychology*, 8, 1681. doi:10.3389/fpsyg.2017.01681
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
- Crossley, S. A., Allen, L. K., Snow, E. L., & McNamara, D. S. (2016). Incorporating learning characteristics into automatic essay scoring models: What individual differences and linguistic features tell us about writing quality. *Journal of Educational Data Mining*, 8(2), 1-19.
- Dweck, C. S., Davidson, W., Nelson, S., & Enna, B. (1978). Sex differences in learned helplessness: II. The Contingencies of evaluative feedback in the classroom and III. An experimental analysis. *Developmental Psychology*, 14(3), 268-276.
- Gates, B. (2007). *Everybody has a robot (in Chinese)* [Scientific People Magazine]. Retrieved from <http://sa.ylib.com/MagArticle.aspx?Unit=featurearticles&id=966>
- Hart, S. A. (2016). Precision education initiative: Moving toward personalized education. *Mind, Brain, and Education*, 10(4), 209-211.
- Herlitz, A., & Rehnman, J. (2008). Sex differences in episodic memory. *Current Directions in Psychological Science*, 17(1), 52–56.
- Hsiao, H. S., Lin, Y. W., Lin, K. Y., Lin, C. Y., Chen, J. H. & Chen, J. C. (2019). Using robot-based practices to develop an activity that incorporated the 6E model to improve elementary school students' learning performances. *Interactive Learning Environments*. doi:10.1080/10494820.2019.1636090.
- Hsiao, J. M., & Huang, Y. C. (2012). An Exploratory study of entrepreneurship in applying LEGO MINDSTORMS® robot on science and creativity education (in Chinese). *Journal of Far East University*, 29(3), 375-386.

- Huang, Y. M., Shadiev, R., & Hwang, W. Y. (2016). Investigating the effectiveness of speech-to-text recognition applications on learning performance and cognitive load. *Computers & Education*, 101, 15-28.
- Hwang, G. J., Sung, H. Y., Chang, S. C., & Huang, X. C. (2020). A Fuzzy expert system-based adaptive learning approach to improving students' learning performances by considering affective and cognitive factors. *Computers & Education: Artificial Intelligence*, 1, 100003. doi:10.1016/j.caeai.2020.100003
- Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Computers & Education: Artificial Intelligence*, 1, 100001. doi:10.1016/j.caeai.2020.100001
- Hwang, G. J., Yang, L. H., & Wang, S. Y. (2013a). A Concept map-embedded educational computer game for improving students' learning performance in natural science courses. *Computers & Education*, 69, 121-130.
- Hwang, M. Y., Hong, J. C., Cheng, H. Y., Peng, Y. C., & Wu, N. C. (2013b). Gender differences in cognitive load and competition anxiety affect 6th grade students' attitude toward playing and intention to play at a sequential or synchronous game. *Computers & Education*, 60(1), 254-263.
- Javora, O., Hannemann, T., Stárková, T., Volná, K. & Brom, C. (2019). Children like it more but don't learn more: Effects of esthetic visual design in educational games. *British Journal of Educational Technology*, 50(4), 1942-1960.
- Kovecses, Z., & Szabo, P. (1996). Idioms: A View from cognitive semantics. *Applied Linguistics*, 17(3), 326-355.
- Krell, M. (2017). Evaluating an instrument to measure mental load and mental effort considering different sources of validity evidence. *Cogent Education*, 4(1), 1280256. doi:10.1080/2331186X.2017.1280256
- Ku, D. T., Huang, Y. H., & Hus, S. C. (2015). The Effects of GBL and learning styles on Chinese idiom by using TUI device. *Journal of Computer Assisted Learning*, 31(6), 505-515.
- Kubilinskiene, S., Zilinskiene, I., Dagiene, V., & Sinkevicius, V. (2017). Applying robotics in school education: A Systematic review. *Baltic Journal of Modern Computing*, 5(1), 50-69.
- Kucuk, S. & Sisman, B. (2017). Behavioral patterns of elementary students and teachers in one-to-one robotics instruction. *Computers & Education*, 111, 31-43.
- Lange, C. & Costley, J. (2019). How sequencing and fading affects the relationship between intrinsic and germane cognitive loads. *Distance Education*, 40(2), 243-261.
- Liao, C. W., Chen, C. H. & Shih, S. J. (2019). The Interactivity of video and collaboration for learning achievement, intrinsic motivation, cognitive load, and behaviour patterns in a digital game-based learning environment. *Computers & Education*, 133, 43-55.
- Liao, Y. W., Zhen, X. J., & Hwang, G. H. (2018, November). *Exploring students' learning outcome and gender differences in a digital video clip course*. Paper presented at the 26th International Conference on Computers in Education, Manila, Philippines.
- Liddy, E.D. (2001). Natural language processing. In *Encyclopedia of Library and Information Science* (2nd ed., pp. 1-15). New York, NY: Marcel Decker, Inc.
- Liu, E. Z. F., & Lin, C. H. (2009). Developing evaluative indicators for educational computer games. *British Journal of Educational Technology*, 40(1), 174-178.
- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An Argument for AI in education*. London, UK: UCL Knowledge Lab.
- Master, A., Cheryan, S., Moscatelli, A., & Meltzoff, A. N. (2017). Programming experience promotes higher STEM motivation among first-grade girls. *Journal of Experimental Child Psychology*, 160, 92-106.
- Mayer, R. E. (2009). *Multimedia learning, second edition*. New York, NY: Cambridge University Press.
- Mutlu-Bayraktar, D., Cosgun, V., & Altan, T. (2019). Cognitive load in multimedia learning environments: A Systematic review. *Computers & Education*, 141, 103618.
- Ng'ambi, D. (2013). Effective and ineffective uses of emerging technologies: Towards a transformative pedagogical model. *British Journal of Educational Technology*, 44(4), 652-661.
- Paas, F. G. W. C., & van Merriënboer, J. J. G. (1993). The Efficiency of instructional conditions: An Approach to combine mental effort and performance measures. *Human Factors*, 35, 737-743.
- Paas, F. G. W. C., & van Merriënboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, 6(4), 351-371.
- Paas, F. G. W. C., Tuovinen, J. E., Tabbers, H., & van Gerven, P. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38(1), 63-71.

- Paas, F. G. W. C., van Gog, T., & Sweller, J. (2010). Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. *Educational Psychology Review*, 22(2), 115–121.
- Papadopoulos, I., Lazzarino, R., Miah, S. & Weaver, T. (2020). A Systematic review of the literature regarding socially assistive robots in pre-tertiary education. *Computers & Education*, 155, 103924.
- Pauls, F., Petermann, F., & Lepach, A. C. (2013). Gender differences in episodic memory and visual working memory including the effects of age. *Memory*, 21(7), 857–874.
- Piaget, J. (1964). *Part I: Cognitive development in children: Piaget development and learning*. New York, NY: Wiley Periodicals, Wiley Company.
- Resnick, M., & Ocko, S. (1991). LEGO/Logo: Learning through and about design. In I. Harel & S. Papert (Eds.), *Constructionism* (pp. 141-158), Norwood, NJ: Ablex Publishing.
- Ring, G. (1991). Student reactions to courseware: Gender differences. *British Journal of Educational Technology*, 22(3), 210-215.
- Rusk, N., Resnick, M., Berg, R., & Pezalla-Granlund, M. (2008). New pathways into robotics: Strategies for broadening participation. *Journal of Science Education and Technology*, 17(1), 59–69.
- Smith, G. G., Haworth, R., & Žitnik, S. (2020). Computer science meets education: Natural language processing for automatic grading of open-ended questions in eBooks. *Journal of Educational Computing Research*, 58(7), 1227-1255.
- Sweller, J. (1998). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257-285.
- Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp.19-30), New York, NY: Cambridge University Press.
- Varney, M. W., Janoudi, A., Aslam, D. M., & Graham, D. (2012). Building young engineers: TASEM for third graders in woodcreek magnet elementary school. *IEEE Transactions on Education*, 55(1), 78–82.
- Webley, P. (1981). Sex differences in home range and cognitive maps in eight-year-old children. *Journal of Environmental Psychology*, 1(4), 293–302.
- Westlund, J. M. K., Jeong, S., Park, H. W., Ronfard, S., Adhikari, A., Harris, P. L., DeSteno, D. & Breazeal, C. L. (2017). Flat vs. expressive storytelling: Young children’s learning and retention of a social robot’s narrative. *Frontiers Human Neuroscience*, 11, 295.
- Wong, L. H., Chin, C. K., & Tan, C. L. (2010). Students’ personal and social meaning making in a Chinese idiom mobile learning environment. *Educational Technology & Society*, 13(4), 15-26.
- Wong, M., Castro-Alonso, J. C., Ayres, P., & Paas, F. (2015). Gender effects when learning manipulative tasks from instructional animations and static presentations. *Educational Technology & Society*, 18(4), 37–52.
- Wu, W. C. V., Wang, R. J., & Chen, N. S. (2015). Instructional design using an in-house built teaching assistant robot to enhance elementary school English-as-a-foreign-language learning. *Interactive Learning Environments*, 23(6), 696-714.
- Yu, G. Z. (1994). *From Xu Xike to Van Gogh (in Chinese)*. Taipei City, Taiwan: Chiu Ko Publishing Co.
- Zhong, B., Zheng, J. & Zhan, Z. (2020). An Exploration of combining virtual and physical robots in robotics education. *Interactive Learning Environments*. doi:10.1080/10494820.2020.1786409



# Factors Affecting the Adoption of AI-Based Applications in Higher Education: An Analysis of Teachers' Perspectives Using Structural Equation Modeling

Youmei Wang<sup>1</sup>, Chenchen Liu<sup>1</sup> and Yun-Fang Tu<sup>1,2\*</sup>

<sup>1</sup>Department of Educational technology, University of Wenzhou, 325035, China // <sup>2</sup>Department of Library and Information Science, Research and Development Center for Physical Education, Health, and Information Technology, Fu Jen Catholic University, Taiwan // wangyoumei@126.com // lcc5424548@126.com // sandy0692@gmail.com

\*Corresponding author

**ABSTRACT:** Owing to the rapid advancements in artificial intelligence (AI) technologies, there has been increasing concern about how to promote the use of AI technologies in school settings to enhance students' learning performance. Teachers' intention to adopt AI tools in their classes plays a crucial role in this regard. Therefore, it is important to explore factors affecting teachers' intention to incorporate AI technologies or applications into course designs in higher education. In this study, a structural equation modeling approach was employed to investigate teachers' continuance intention to teach with AI. In the proposed model, 10 hypotheses regarding anxiety (AN), self-efficacy (SE), attitude towards AI (ATU), perceived ease of use (PEU) and perceived usefulness (PU) were tested, and this study explored how these factors worked together to influence teachers' continuance intention. A total of 311 teachers in higher education participated in the study. Based on the SEM analytical results and the research model, the five endogenous constructs of PU, PEU, SN, and ATU explained 70.4% of the changes in BI. In this model, SN and PEU were the determining factors of BI. The total effect of ATU was 0.793, followed by SE, with a total effect of 0.554. As a result, the intentions of teachers to learn to use AI-based applications in their teaching can be predicted by ATU, SE, PEU, PU and AN. Among them, teachers' SE positively influenced teachers' PEU and ATU towards adopting AI-based applications, and also influenced PU through PEU. In addition, the relationship between teachers' SE and AN was negatively correlated, which indicated that enhancing teachers' SE could reduce their AN towards using AI-based applications in their teaching. Accordingly, implications and suggestions for researchers and school teachers are provided.

**Keywords:** Artificial intelligence, Higher education, Anxiety, Self-efficacy, Technology acceptance model

## 1. Introduction

With their development, technologies have had substantial influences on teaching management, teaching innovation and the analysis of learning behavior (Nelson et al., 2019). In particular, the development of speech recognition, natural language recognition and deep learning has fostered educators' attention to artificial intelligence (AI) technologies. AI has been described as computers being used to mimic human minds so as to perform cognitive tasks (e.g., thinking, learning, problem solving) (Hwang, 2003; Nilsson, 2014). AI technologies can analyze learners' learning process, provide adaptive learning resources, and provide evaluation and suggestions based on learners' performance, which can serve as a learning diagnostic tool (Colchester et al., 2017; Hwang et al., 2011; Timms, 2016).

AI technologies have been gradually changing the role of teachers in learning activities; teachers can select appropriate AI teaching tools to monitor learners' learning processes and offer them personalized and timely assistance (Edwards et al., 2018). Researchers have indicated that developing a virtual laboratory, an intelligence teaching platform or an intelligence learning tool based on AI technologies can support diverse learning approaches, provide learners with personalized guidance, learning prompts and feedback, and assist learners in developing higher order thinking abilities as well (Hwang, 2014; Lin et al., 2018; McArthur et al., 2005). Moreover, with the development of communication and computing technologies, Artificial Intelligence in Education (AIED) has become an important issue in education (Hwang et al., 2020c; Chen et al., 2020b).

From the perspective of precision education, AI technologies could analyze and predict learners' academic achievement, and intelligent tutoring systems (ITSSs) could provide personalized instruction or support to students by understanding learners' learning status and behavior, diagnosing students' learning status and giving feedback automatically, to assist teachers with instructional assessment (Chen et al., 2020a; Hwang et al., 2020c; Hwang et al., 2014; Lin et al., 2021). AIED is a highly technology-dependent and cross-disciplinary field, and while AI technologies are being integrated into education, their use in teaching remains a challenge; for example, researchers might fail to effectively implement AIED applications and activities without understanding the role of

AI in education and the functions of AI technologies (Hwang et al., 2020a). In addition, teachers who understand the functions and attributes of AI technologies can adopt suitable AI applications in their classrooms to promote students' motivation, engagement, or learning achievement (Chen et al., 2020a; Hwang et al., 2020c; Hwang et al., 2021). In this situation, it is crucial to understand teachers' standpoints on employing AI in teaching (e.g., their attitude towards AI and intention to use it) because teachers' acceptance or rejection will affect the application of AI to the teaching process (Popenici & Kerr, 2017).

Teacher acceptance has been proven to be an essential element in the process of educational innovation (Chen et al., 2009; Sánchez-Prieto et al., 2017). For instance, some studies have explored teachers' acceptance of adopting mobile technologies or digital technologies in teaching activities, while others have examined teachers' self-efficacy, perceptions (including usefulness and ease of use), feelings and attitudes towards adopting technologies (Nikou & Economides, 2017; Sánchez-Prieto et al., 2016; Scherer et al., 2019; Teo et al., 2008). Researchers have indicated that teachers' attitudes towards the adoption of AI technologies determine whether they will be used to support teaching activities, and the degree to which the technologies and actual teaching practice are integrated (Becker et al., 2017; Edwards et al., 2018; Wang & Wang, 2009).

In the field of education research, the Technology Acceptance Model (TAM) is most commonly used to explain teachers' attitudes and behavioral intention to use novel technologies to support teaching activities (Al-Emran et al., 2018; Scherer et al., 2019; Teo, 2019). On the other hand, researchers have pointed out the extra work that teachers need to do to prepare the new materials or to start teaching activities for the new technology/system, the time it takes to perform the necessary training, and the anxiety that comes from not being able to smoothly use the new technology/equipment (Sánchez-Prieto et al., 2017). Studies have also specified that reducing teachers' anxiety about the adoption of technologies and promoting teachers to effectively apply technologies in class can strengthen their confidence in adopting technologies (Clark-Gordon et al., 2019; Lim & Khine, 2006; Sánchez-Prieto et al., 2017). Sánchez-Prieto et al. (2017) also reported that teachers' beliefs about their ability to perform their tasks and achieve their goals were stronger in facilitating attitudes and willingness to adopt technology in teaching. Teachers' adoption of technology/systems in their teaching is a complex and multi-directional issue, and if teachers lack sufficient motivation and intention to employ technology/systems, then the unused technology/systems will eventually become useless (Bai et al., 2019; Hwang et al., 2020a; Teo, 2019; Sánchez-Prieto et al., 2016; Wang & Wang, 2009). Therefore, the present study aimed to investigate teachers' attitudes towards and intentions to adopt AI-based applications in their teaching, and based on TAM with the extension of two constructs: anxiety and self-efficacy, to explore teachers' perspectives, attitudes and behavioral intentions to integrate AI-based applications into teaching. The findings could be a good reference for those instructors and policymakers in schools or institutes.

## **2. Literature review and model development**

### **2.1. Artificial intelligence in education (AIED)**

Due to the advancements in computer technology, the development of computer systems that are closer to human reasoning, decision-making and problem-solving capabilities has also received increasing attention. AI aims for human-level intelligence; researchers define AI as a computer-controlled device which has a human-like manner and is able to perform tasks such as learning, reasoning and self-correction (Chen et al., 2020b; Hwang, 2003; Nilsson, 2014; Shi & Zheng, 2006). Also, AI is referred to as Machine Intelligence or Computational Intelligence. In the past decades, researchers have attempted to apply AI to different fields such as playing chess, speech recognition, writing poetry, Intelligent Personal Assistants (IPAs) and diagnosing diseases (Aibinu et al., 2012; Hwang et al., 2020c; Russell & Norvig, 2003).

AIED has become one of the current emerging fields of novel educational technology. AI technologies overcome the limitations of space and time; with the portability of mobile devices, learners can read the materials, practice and collect information at any time. In the meantime, AI learning systems can provide learning guidance and required auxiliary materials based on the learners' environment (Hung et al., 2014; Liu et al., 2019). Zawacki-Richter et al. (2019) reviewed the papers relevant to AIED published from 2007 to 2018, and found that the main application fields of AIED were profiling and prediction, assessment and evaluation, adaptive systems and personalization, and ITSs. For instance, ITSs can provide personalized learning interfaces and materials by analyzing students' personal learning characteristics and status (Chen et al., 2020a). Also, it can select teaching strategies and approaches based on students' current status and provide students with adequate assistance and timely guidance in order to facilitate the effectiveness of learning (Huang & Chen, 2016; Hwang, 2003; Van Seters et al., 2012). Moreover, in adaptive and intelligent web-based educational systems, taking into account

both the affective and cognitive status of individual learners, the adaptive learning model could improve learners' learning outcomes and assist low achievers in successfully completing learning tasks (Hwang et al., 2020b). Some scholars have also tried to build user learning models by targeting large-scale data sources in learning systems and educational environments with big data analysis (e.g., Rau et al., 2017).

The interaction data analyzed by AIED to support learners' learning processes can serve as a mentor for every learner. Besides, AIED can provide insights into students' learning progress so that teachers can actively offer support and guidance when students are in need (Hwang et al., 2020c; Hwang et al., 2021; Woolf et al., 2013). However, researchers have indicated that applying technologies in educational environments should consider learning content, pedagogy and the environment created by the students, teachers and technology (Hsieh & Tsai, 2017; Oblinger, 2012). Some researchers have also found that teachers' acceptance level of AI technologies will influence the integration of AI and teaching activities, which is also one of the challenges of AIED (Ifinedo et al., 2020; Popenici & Kerr, 2017; Teo et al., 2008; Zawacki-Richter et al., 2019). As a result, understanding teachers' acceptance of AI and relevant influencing factors is a current important research issue.

## **2.2. Technology acceptance model (TAM)**

The TAM was first proposed by Davis et al. (1989) to explore users' acceptance of technologies. TAM emphasizes the users' intention to use or their actual use of technologies (Al-Emran et al., 2018; Bai et al., 2019; Legris et al., 2003). When users believe that technologies are helpful, they will then adopt those technologies and have a positive attitude towards them. On the other hand, when users think that specific technologies are easy to use and can help them complete tasks more effectively, they generally have stronger willingness to adopt them (Davis, 1989; Sánchez-Prieto et al., 2017; Teo, 2019; Wang & Wang, 2009). In other words, if the technologies are not easy to use, users will maintain the status quo or choose other options even if the technology is helpful (Teo, 2019). Studies have also indicated the importance of teachers' attitudes towards the integration of new technologies (including mobile learning platforms, virtual environments) into teaching for their adoption behavior (Dávideková et al., 2017; Hsieh & Tsai, 2017; Ifinedo et al., 2020).

Several studies have adopted TAM to explain teachers' intentions and behavior of employing new technologies in their teaching activities (Al-Emran et al., 2018; Scherer et al., 2019; Teo, 2019). For instance, teachers' self-efficacy for new technologies will influence the positive evaluation of their perceptions (e.g., perceived usefulness and ease of use), which will then affect their attitude and behavior of using new technologies when teaching. Other studies have specified that teachers' positive or negative perceptions when adopting new technologies will also affect their attitude and behavior of adoption (Bai et al., 2019; Sánchez-Prieto et al., 2017). Besides, researchers have specified that after users employ the technology, as they become familiar with the technology, their concern about "ease of use" becomes less, which could influence users' perceptions of its ease of use as well as their attitude toward their adoption of the technology (Lin, 2011; Teo, 2019; Wang & Wang, 2009). With the development of technologies, constructing smart learning environments (SLEs) to support teaching and learning has become a trend and a crucial goal for educational practitioners. This highlights that teachers play an important role in the process of applying AI technologies in teaching and learning activities (Kinshuk et al., 2016; Hwang, 2014). As a result, based on TAM, the present study examined teachers' perspectives, attitude and behavioral intention to integrate AI technologies into teaching.

## **2.3. Self-efficacy (SE)**

In the context of information technology, SE is often defined as one's SE of using that technology, which refers to one's own judgement about one's ability to complete a specific task by using technology (Compeau & Higgins, 1995; Teo, 2019). Some studies have indicated that SE not only directly influences users' perceived usefulness of the technology, but also affects their attitudes towards the adoption of the technology (Motaghian et al., 2013; Teo & Zhou, 2014; Yeşilyurt et al., 2016). Teachers' SE is defined as their belief in their own capabilities. This can facilitate students' learning and is also the key point of integrating technology into teaching (van Dinther et al., 2013). Researchers have found that teachers with higher SE were more likely to successfully integrate teaching into their instruction (Bai et al., 2019). For example, in the flipped teaching activities in class, university instructors' SE influences their attitude towards using technology (Lai et al., 2018). The abovementioned studies indicated that teachers' SE in technologies is the belief in applying technologies when teaching, which has effects on their ease of use and attitude (Teo & Zhou, 2014; Yeşilyurt et al., 2016).

## 2.4. Anxiety (AN)

AN is generated due to users' anxious and nervous feelings about novel technologies. Studies have specified that users' negative feelings caused by the adoption of new technologies, such as AN, might negatively influence their attitude and SE (Agudo-Peregrina et al., 2014; Cazan et al., 2016). The relationship between AN and users' adoption of new technologies has been verified, for example, anxiety has a negative effect on teachers' and students' attitude towards the adoption of mobile technologies (MacCallum & Jeffrey, 2014). Studies have revealed that university teachers' attitudes towards adopting technologies when teaching are influenced by their AN. That is to say, teachers' feelings (positive or negative) about integrating technologies into teaching affects their adoption attitude (Clark-Gordon et al., 2019; Park et al., 2019).

## 2.5. Research model and hypotheses

Since Davis proposed the TAM model, it has been extensively verified and applied by the industry and academia in numerous relevant studies. Especially for teachers who integrate technologies into teaching, it also has its predictive power (Ifenthaler & Schweinbenz, 2013; Sánchez-Prieto et al., 2017; Teo, 2019; Ursavaş et al., 2019; Wang & Wang, 2009). Based on TAM, the present study adopted the six factors of AN, SE, PU, PEU, ATU and BI to explore teachers' perspectives, attitude and behavioral intention to employ AI-based applications to support their teaching. The research model is shown in Figure 1.

According to the literature, the university teachers' PEU, PU, and attitudes towards adopting technologies for teaching could have effects on their BI; their PEU and PU also influence their attitudes toward adopting AI applications in teaching activities (Kao & Tsai, 2009; Teo, 2019; Wang & Wang, 2009). Also, university teachers' PEU and PU of adopting technologies could directly or indirectly affect their BI. Researchers have also shown that teachers' PEU of using AI applications could affect perceptions of PU, ATU, and BI (Kao & Tsai, 2009; Wang & Wang, 2009). Therefore, based on TAM, the present study investigated university teachers' acceptance of AI technologies and relevant influencing factors. The research hypotheses of the present study are as follows:

H1: PU has a significant positive effect on ATU.

H2: PEU has a significant positive effect on PU.

H3: PEU has a significant positive effect on ATU.

H8: PU has a significant positive effect on BI.

H9: ATU has a significant positive effect on BI.

H10: PEU has a significant positive effect on BI.

In this study, SE refers to the measure or extent of university teachers' beliefs about the integration of using technologies in their teaching activities. Previous studies have shown that SE as an individual factor in explaining university teachers' beliefs of using technologies in teaching directly affects their PEU and attitudes toward technology adoption (Kao & Tsai, 2009; Ursavaş et al., 2019; Wang & Wang, 2009). A higher degree of SE implies a greater degree of perceived PEU and ATU, which may lead to use of AI-based applications for teaching. Accordingly, the following research hypotheses are proposed:

H4: SE has a significant positive effect on PEU.

H6: SE has a significant positive effect on ATU.

Moreover, researchers have also pointed out that SE directly links to AN (Kao & Tsai, 2009; Sánchez-Prieto et al., 2017). Some studies have implied that when teachers lack the ability to use new technologies, they may have negative perceptions of the technologies (e.g., anxiety). This could influence their cognition of the functions and their attitude towards using the technologies, which must be guided and assisted by teacher training (Cheok et al., 2017; Sánchez-Prieto et al., 2017). When teachers are more familiar with or more confident in using the technologies, they may find that it is easier to use them to assist with their teaching; on the contrary, if teachers experience frustration or negative feelings, it may then influence their attitude towards the adoption of technologies (Motaghian et al., 2013; Sánchez-Prieto et al., 2017; Wang & Wang, 2009). Thus, the following hypotheses are proposed:

H5: SE has a significant negative effect on AN.

H7: AN has a significant negative effect on ATU.

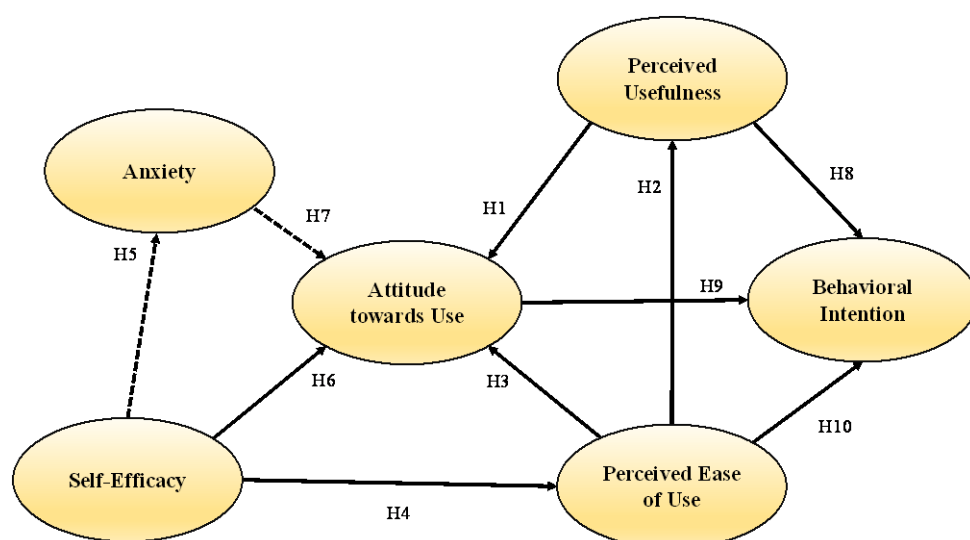


Figure 1. Proposed research model. Note. — positive effect, ---- negative effect

### 3. Method

#### 3.1. The participants

The participants in the present study were university in-service teachers in China. They had the same experience of using AI technologies (e.g., Mosoteach, Smart Class, Youdao Translation, HappyClass Smart Classroom System) and had received the same training courses. The study collected a total of 311 valid questionnaires from 171 male and 140 female participants by excluding the teachers who had no teaching experience or no Internet usage experience as well as the questionnaires with incomplete answers in February and March, 2020. As for age, 45.66% were 30 to 40 years old; 36.01% were 40 to 50 years old; 10.93% were above 50 years old; and 7.40% were under 30 years old.

Figure 2 shows one of the intelligent applications, “Mosoteach” designed for English teachers. The app helps teachers to collect students’ work efficiently, and also helps teachers to save the time spent correcting assignments manually in previous teaching contexts. In addition, the app can provide targeted advice to students on how to improve their English writing in detail, which also saves teachers’ communication time with each student.



Figure 2. An example of AI-based applications adopted

### 3.2. Instruments

The present study referred to Davis (1989) and adopted scale items from the previous studies. These items were adapted from published sources that reported a high degree of reliability (Sánchez-Prieto et al., 2017; Teo, 2019; Ursavaş et al., 2019; Wang & Wang, 2009). The instrument includes participants' demographic information and 21 items. These items aim to evaluate participants' belief in the following constructs: PU (four items: e.g., "I think it is useful to learn to adopt AI tools to support teaching"), PEU (three items: e.g., "For me, learning to operate AI tools to support teaching activities is easy"), SE (three items: e.g., "I have the skills required to use AI tools to support teaching"), AN (four items: e.g., "I think it is very difficult to use AI tools to support teaching activities"), ATU (three items: e.g., "For me, learning to operate AI tools to support teaching activities is easy"), BI (four items: e.g., "I will actively learn to adopt AI tools to assist in teaching"). These items were adapted from published sources that reported a high degree of reliability (Sánchez-Prieto et al., 2017; Teo, 2019; Ursavaş et al., 2019; Wang & Wang, 2009).

In order to make the questionnaire content in accordance with the teachers' experience of using AI-based applications in teaching contexts, the present study consulted two professors who are experts in the AI field and two experts who are familiar with the integration of technology into teaching. They helped confirm that all items in the questionnaire were in line with teachers' familiar tone of expression, and could be used to realize teachers' perceptions of and attitudes towards AI tool-supported teaching as a reference for future university teachers to promote AI tools to support teaching activities. The questionnaire in the study adopted a 5-point Likert scale, ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). The preliminary analysis indicated that the factor loadings of four items (i.e., PU4, PEO3, SE3 and ATU3) were lower or had a high correlation with other items in the model. As a result, these items were removed from further analysis; a total of 17 items were used for the following analysis (Appendix A). The final structure showed good internal consistency, reliability, and Cronbach's alpha values; the Cronbach's alpha values are listed in Table 1, and range from .699 to .925.

### 3.3. Data analysis

The present study employed AMOS in SPSS for the analysis. First of all, the descriptive statistics were conducted to verify the skewness and kurtosis of values and to establish the univariate normality of the data. The critical values were  $\pm 3.0$  and  $\pm 10.0$ , respectively (Kline, 2010). Furthermore, researchers tested the multivariate normality using Mardia's normalized multivariate kurtosis (Mardia, 1970). Afterwards, confirmatory factor analysis (CFA) was performed to examine the structural validity of the questionnaire. Finally, we verified the path model hypothesized to examine the effects of the influences of university teachers on PU, PEU, SE, AN, ATU and BI of adopting AI tools.

## 4. Results

### 4.1. Descriptive statistics

The means, SDs, skewness, and kurtosis values for each of the 17 items in the questionnaire were computed. The mean and standard deviation of AN were 2.842 and .899, respectively. The means of the other constructs were between 3.982 and 4.092 with standard deviations between .550 and .674. This represented participants' positive responses to the items and the mean of values distribution. The values of the skewness and kurtosis for the items were between -1.082 and .427, and -.781 and 3.385, respectively. These values were within the recommended cutoffs of  $\pm 3.0$  and  $\pm 10.0$  for skewness and kurtosis, respectively, indicating univariate normality in the data (Kline, 2010). Finally, Mardia's multivariate kurtosis value was calculated as 133.350 and using the Raykov and Marcoulides (2008) formula,  $p(p+2)$  was calculated as 323. Since the multivariate kurtosis value was smaller than 323, the data showed multivariate normality.

### 4.2. Test of the measurement model

The present study applied the CFA evaluation model, including the six constructs of PU, PEU, SE, AN, ATU and BI (see Figure 4). The overall model fit evaluation adopted  $\chi^2$  and other fit indices such as the Tucker-Lewis index (TLI), the comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR). Hu and Bentler (1999) pointed out that the TLI and CFI values were higher than 0.95, which indicated a good model fit. Also, it was acceptable that RMSEA and SRMR were

lower than 0.06 and 0.08, respectively. From the results, the measurement model displayed an acceptable fit to the sample data ( $\chi^2 = 194.48$ ;  $\chi^2/df = 1.870$ ; TLI = .967; CFI = .975; RMSEA = .053; SRMR = .037).

Table 1 presents the CFA results; all the factor loadings of the measured items were higher than the threshold value of .60 (ranging from .711 to .922). The Cronbach's alpha values of PU, PEU, SE, AN, ATU and BI were .843, .899, .887, .925, .699 and .916, respectively. The overall reliability of the questionnaire was .809, indicating sufficient internal consistency of the factor items. Moreover, the range of composite reliability (CR) was .719~.925, and the range of average variance extracted (AVE) was .562~.818, indicating that the present study had good convergence validity of the adopted variables. The convergence validity of the variables in the present study all meet the standard (Fornell & Larcker, 1981).

In addition to convergence validity, the square roots of all the AVEs in the present study were greater than their correlation coefficients; therefore, each variable adopted in the study had its discriminant validity (Farrell, 2010) (see Table 2).

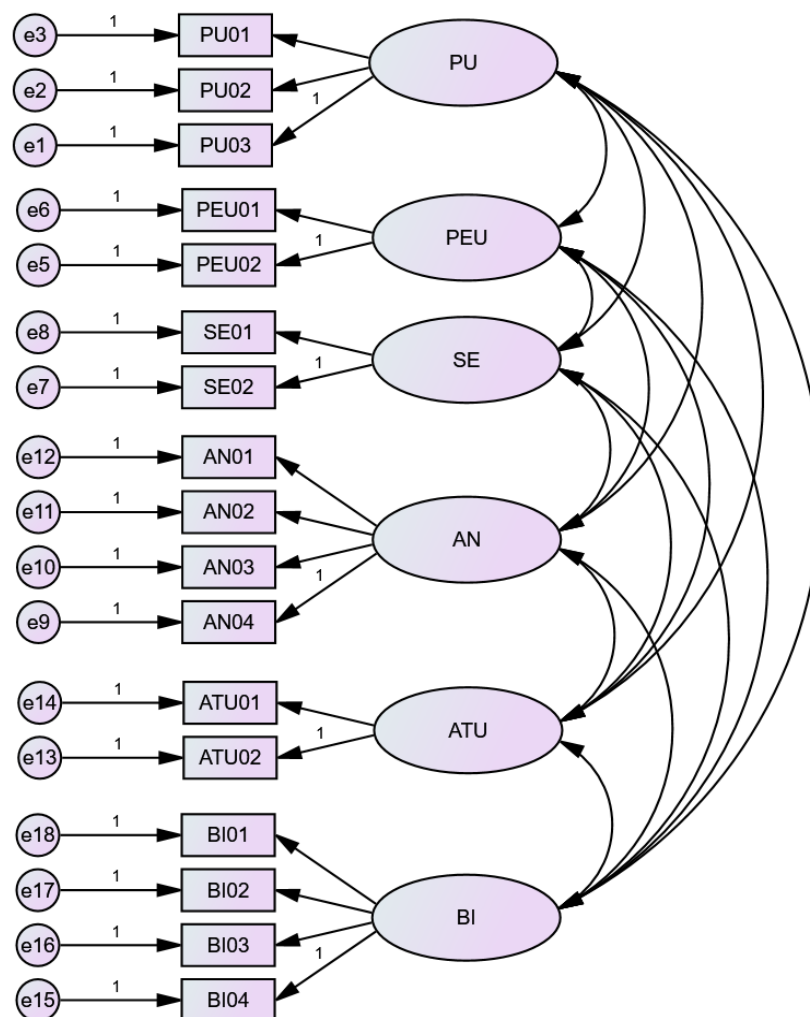


Figure 3. Measurement model with 17 items.

Table 1. Results of the CFA

Items	UE	t-value	SE	CR	AVE	Alpha value	Mean	SD
PU				0.845	0.646	0.843	4.070	0.636
PU01	1.016	13.134	0.789					
PU02	1.239	13.6	0.868					
PU03 <sup>#</sup>	1.000		0.749					
PEU				0.900	0.818	0.899	4.0482	0.674
PEU01	0.925	17.804	0.886					
PEU02 <sup>#</sup>	1.000		0.922					
SE				0.886	0.796	0.887	3.9823	0.647

SE01	0.969	17.476	0.879					
SE02 <sup>#</sup>	1.000		0.905					
AN				0.925	0.755	0.925	2.842	0.899
AN01	0.957	19.37	0.853					
AN02	1.046	20.819	0.888					
AN03	1.019	20.178	0.873					
AN04 <sup>#</sup>	1.000		0.862					
ATU				0.719	0.562	0.699	4.092	0.623
ATU01	0.679	11.524	0.711					
ATU02 <sup>#</sup>	1.000		0.787					
BI				0.916	0.732	0.916	4.036	0.550
BI01	1.043	19.633	0.851					
BI02	1.021	19.409	0.854					
BI03	1.035	19.459	0.849					
BI04 <sup>#</sup>	1.000		0.868					

Note: UE= unstandardized estimate; SE= standardized estimate, factor loadings; PU= perceived usefulness; PEU= perceived ease of use; SE= self-efficacy; AN= anxiety; ATU= attitude towards use; BI= behavioral intention. \*  $p < .01$ ; <sup>#</sup> this value was fixed at 1.000 for model identification purposes.

Table 2. Correlation coefficient and discriminant validity

	BI	ATU	AN	SE	PEU	PU
BI	(0.856)					
ATU	0.823	(0.750)				
AN	-0.128	-0.307	(0.869)			
SE	0.596	0.666	-0.200	(0.892)		
PEU	0.524	0.640	-0.208	0.654	(0.904)	
PU	0.495	0.439	-0.111	0.426	0.460	(0.804)

Note. Diagonal values shows square root of AVE; PU = perceived usefulness; PEU = perceived ease of use; SE = self-efficacy; AN = anxiety; ATU = attitude towards use; BI = behavioral intention.

#### 4.3. Tests of direct and indirect effects

Results of the test of the structural model showed a good model fit ( $\chi^2 = 212.298$ ;  $\chi^2/df = 1.948$ ; TLI= 0.964; CFI= 0.971; RMSEA= .055; SRMR= .048). From the research model (Figure 1), four endogenous constructs were tested. Based on the hypotheses proposed in this study, the bootstrap method was performed for the evaluation. As shown in Table 3, seven out of 10 hypotheses were supported by the data; except for H1, H7 and H10, all the hypotheses were supported in the present study.

Table 3. Hypothesis Testing Results

Hypotheses	Path	Estimate	t-value	Bias-corrected		Sig <i>p</i>	Result
				Lower	Upper		
H1	PU→ATU	0.125	1.81	-0.043	0.267	0.146	Not supported
H2	PEU→PU	0.472	7.254	0.321	0.615	0.002	Supported
H3	PEU→ATU	0.279	3.119	0.065	0.508	0.014	Supported
H4	SE→PEU	0.663	11.596	0.566	0.748	0.002	Supported
H5	SE→AN	-0.209	-3.367	-0.339	-0.071	0.003	Supported
H6	SE→ATU	0.437	5.503	0.246	0.611	0.001	Supported
H7	AN→ATU	-0.089	-1.623	-0.202	0.016	0.107	Not supported
H8	PU→BI	0.175	2.881	0.039	0.359	0.012	Supported
H9	ATU→BI	0.793	7.828	0.626	1.003	0.001	Supported
H10	PEU→BI	-0.058	-0.734	-0.297	0.132	0.533	Not supported

Note. PU = perceived usefulness; PEU = perceived ease of use; SE = self-efficacy; AN = anxiety; ATU = attitude towards use; BI = behavioral intention.

Table 4 shows the standardized total effects, as well as the direct and indirect effects of each variable correlated in the model. The sum of direct and indirect effects is total effects. In the model of the present study, the standardized total effects of predictor variables on the dependent variables was between -.209 and .793.



Four endogenous constructs were tested in the model. The coefficient of variation of BI was determined by PU, PEU and ATU, and the explanatory power ( $R^2$ ) was .704. In other words, AN, SE, PU, PEU and ATU jointly explained 70.4% of BI changes. The most dominant determinant was ATU with a total effect of .793, followed by SE with a total effect of .554, PEU with a total effect of .292, PU with a total effect of .274, and AN with a total effect of -.071.

Among these four endogenous constructs, the highest amount of variance (54.5%) was explained by the determinants of ATU. The most dominant determinant was SE with a total effect of .679, followed by PEU with a total effect of .338, PU with a total effect of .125, and AN with a total effect of -.089. The explained variation of PEU in this model was 43.9%; the most dominant determinant was SE with a total effect of .663. The explained variation of PU was 22.3%; the most dominant determinants were SE and PEU with a total effect of .313 and .472, respectively. The explained variation of AN was 4.4%; the most dominant determinant was SE with a total effect of -.209.

*Table 4. Direct, indirect and total effects of the research model*

Endogenous variable	Determinant	Standardized estimates		
		Direct	Indirect	Total
AN ( $R^2 = 0.044$ )	SE	-0.209	-	-0.209
PU ( $R^2 = 0.223$ )	SE	-	0.313	0.313
	PEU	0.472	-	0.472
PEU ( $R^2 = 0.439$ )	SE	0.663	-	0.663
ATU ( $R^2 = 0.545$ )	AN	-0.089	-	-0.089
	SE	0.437	0.242	0.679
	PU	0.125	-	0.125
	PEU	0.279	0.059	0.338
	BI	-	-0.071	-0.071
BI ( $R^2 = 0.704$ )	SE	-	0.554	0.554
	PU	0.175	0.099	0.274
	PEU	-0.058	0.350	0.292
	ATU	0.793	-	0.793

*Note.* PU = perceived usefulness; PEU = perceived ease of use; SE = self-efficacy; AN = anxiety; ATU = attitude towards use; BI = behavioral intention.

## 5. Discussion and conclusion

The present study was based on TAM and added teachers' SE and AN about integrating AI tools to examine university teachers' perspectives on AI tool-supported teaching as well as their behavior and influencing factors. Besides, the research model was tested, in which individual differences such as technology AN, SE and relevant factors affecting teachers' acceptance of technology were discussed.

The findings of this study highlight that teachers' SE would positively influence their PEU and ATU about adopting AI technologies, and it could further affect PU through PEU. This is in line with Agudo-Peregrina's et al. study (2014) which revealed the dual nature of perceived usefulness: the component related to efficiency and performance and the component related to flexibility. For instance, teachers would discover that there were differences in efficiency and performance-related advantages of AI tools, and they would also consider the high correlation between the chosen learning strategy and the academic performance (Agudo-Peregrina et al., 2014; Paechter et al., 2010). On the other hand, Bai et al. (2019) illustrated that teachers' SE usually has an indirect effect on their attitude to adopt certain technology in teaching. Chang et al. (2017) also found that the relationship between SE and PU is influenced by PEU. In other words, university teachers' SE would have positive effects on their perceived ease of use, perceived usefulness and attitude toward AI technologies. AI technology for teaching is still in its early stage; thus, most of the teachers still worry whether their ICT skills could meet the needs of integrating artificial intelligence in teaching practice. During the training process, it is necessary to increase their ability and confidence in learning to adopt AI tools, thus making them feel that it is easy to apply them in their teaching. On the other hand, teachers' confidence in using AI technology makes them feel that they have control in the teaching environment, and as such, the application of AI technology is not complicated for them, so they can easily integrate it into their teaching activities

Moreover, teachers' SE and AN were negatively correlated, denoting that teachers with higher SE were less anxious about integrating AI technologies into their teaching (Yeşilyurt et al., 2016). Bai et al. (2019) employed the technology acceptance model, the value-expectancy theory and a learning perspective to discuss the effects of teacher professional development. Researchers have indicated that teachers' ICT self-efficacy would positively affect their continuance intention through their perceptions (i.e., perceived ease of use and perceived usefulness). Also, teachers' ICT anxiety would have negative impacts on their perceived ease of use and continuance intention. Researchers have also reported that anxiety is related to prior unpleasant experiences, and therefore, anxiety could potentially neutralize the effects of PEU (Chavoshi & Hamidi, 2019). In particular, in China, the examination-oriented culture may be an explanation, since most teachers face a heavy workload and they may be concerned with the overtime spent on learning new technology. From the perspective of facilitating the AI technology acceptance of future teachers, it is important to design educational actions that emphasize the usefulness of these AI technologies in teaching and learning practice, and reduce the anxiety they may generate. These points should be taken into account when planning teacher training, which should focus on the pedagogical use of these AI technologies in real teaching and learning environments through the practical activities (Bai et al., 2019; Sánchez-Prieto et al., 2016; Sánchez-Prieto et al., 2017).

Another finding is that the teachers' PEU would positively affect their PU as well as their attitude toward applying AI technologies to support teaching. The findings were consistent with the interaction relationships between PEU, PU and ATU of the technology acceptance research in the educational field (Teo, 2019; Joo et al., 2018). For example, Teo (2019) pointed out that the interaction between PU, PEU, FC and subjective norms had influences on ATU, which then facilitated teachers' intention to use technology. University teachers' perceived ease of use of AI technologies would directly influence the perceived usefulness, and the perceived ease of use of AI technologies had a significant influence on teachers' adoption of AI technologies in teaching. The present study also uncovered that teachers' perceived usefulness of AI technologies and their attitude towards AI technology-supported teaching would have positive effects on their adoption behavior. For instance, Sánchez-Prieto et al. (2017) examined the differences of acceptance between higher education and lifelong learning on the digital learning system, and suggested building up stronger relationships between perceived usefulness and behavioral intention, perceived ease of use and perceived usefulness as well as SE and perceived ease of use.

Some of the hypotheses in this study are not significant. For instance, PU did not have a significant influence on ATU (H1). Raza et al. (2017) had a similar finding of the insignificant impact of PU on ATT to adopt mobile banking. A possible explanation is that university teachers tend to insist on their point of view based on their own experience; thus, PU may easily affect the behavior intention rather than attitude. Besides, AN did not decide their attitude towards use (H7). It is indicated that AN often negatively impacts ATT or BI in the context of education (Hsu, 2009). The significance of the effect of AN on ATT may depend on which situation causes teachers' negative feelings, discomfort or reluctance to adopt AI technologies: subject matters or SE for ICT skills. Thus, introducing a few carefully designed supportive activities in teachers' training programs may help familiarize the teachers with AI technology and raise their comfort levels. Contrary to our expectations, PEU did not have a significant influence on BI (H10). In other words, if AI technologies are not easy to use and apply, even those that are useful for teachers and learners, teachers may remain in their original status or choose other options (Ursavaş et al., 2019). The participants of the current study who had experience of integrating technologies into their teaching practice tended to focus more on perceived usefulness for teaching and learning. In other words, even if some AI technologies are easy to use and apply, without improving the quality of teachers' instruction, it would not significantly change teachers' behavior when adopting these technologies in their teaching activities.

AI technologies can analyze students' learning behavior and performance and provide students with just-in-time guidance and feedback. Moreover, they can also integrate students' individual and learning process data, diagnose students' learning situation, and assist teachers in adjusting the teaching strategies, which then enhance students' learning effectiveness (Hwang, 2014; Hung et al., 2014). The findings of this study specified that university teachers' adoption of AI technologies in their teaching would be influenced by their perceived usefulness and attitude towards AI technologies, for example, how to effectively increase students' learning effectiveness through AI technologies (Hung et al., 2014). One possible explanation is that the information skills of teachers now have a certain degree of training basics; when teachers consider integrating technologies into teaching, they directly take the usefulness of technologies for teaching into consideration, and evaluate whether to adopt or keep employing them (Wang & Wang, 2009).

On the other hand, the ease with which university teachers adopt AI technologies also affects their attitude towards using AI to support teaching. Besides, teachers' perceived ease of use of AI further influences their perceived usefulness as well as their behavior of employing AI to support teaching. In other words, increasing teachers' ease of integrating AI technologies into teaching activities can also enhance their perceived usefulness

of AI technology-assisted teaching, and facilitate their adoption behavior. Aside from improving the user interface for using AI technologies, some studies have pointed out that teachers' confidence and ability of using AI technologies could affect their willingness of incorporating AI technologies into their learning designs (Sánchez-Prieto et al., 2017). Based on the research results, educational implications for teacher education in higher education were concluded. Firstly, the results of the study informed the educators and policymakers in higher education that when planning training activities of adopting AI to support teaching for teachers, it is necessary to consider teachers' individual differences and determine effective ways to mitigate teachers' AN or strengthen their SE of adopting AI technologies in teaching. For instance, enhancing teachers' professional development through teacher training or assistance from technology professionals can help teachers spend less time learning how to adopt AI technologies in their teaching (Cheok et al., 2017; Kao & Tsai, 2009; Wang & Wang, 2009). Besides, the use of AI technologies has spread to every corner of modern society; it is therefore necessary to inform teachers who may have diverse educational backgrounds of the basic concepts of Artificial Intelligence and provide convenient AI tools for teachers to integrate into their teaching processes.

The present study has some limitations. In terms of samples, the participants were recruited from among university teachers in China, which may limit the research inference. It is suggested that researchers can further investigate the factors affecting intentions of teachers with different backgrounds and teaching experiences to use AI technologies in school settings in the future. Regarding university teachers' attitudes towards and behavior of adopting AI technologies to support teaching, some external variables can be considered such as social support, subjective norms, and facilitating conditions, to name just a few. Furthermore, future researchers can collect and compare the data from different points in time to understand the effects of the evolution of teachers' attitudes towards AI-supported teaching. Also, future studies can design the intervention experiment and interviews to explore the implementing strategies and application effects of a mixed teaching approach based on a smart learning environment in different teaching contexts to obtain a deeper understanding of its influences on teachers' attitudes towards and perspectives on AI-supported teaching. Moreover, the transformation of the role of university teachers in AI technology-integrated teaching and learning activities (e.g., collaborative learning facilitator, learning evaluator, feedback giver) is also an issue that is worth investigating. Future studies need not only rely on technologies, algorithms and teaching strategies, but should also focus on teachers' adoption attitude toward AI technologies as well as their practice of applying AI in their teaching, which creates a meaningful learning environment.

## References

- Agudo-Peregrina, Á. F., Hernández-García, Á., & Pascual-Miguel, F. J. (2014). Behavioral intention, use behavior and the acceptance of electronic learning systems: Differences between higher education and lifelong learning. *Computers in Human Behavior*, 34, 301–314.
- Aibinu, A. M., Salami, M. J. E., & Shafie, A. A. (2012). Artificial neural network based autoregressive modeling technique with application in voice activity detection. *Engineering Applications of Artificial Intelligence*, 25(6), 1265–1276.
- Al-Emran, M., Mezhyuev, V., & Kamaludin, A. (2018). Technology acceptance model in M-learning context: A Systematic review. *Computers & Education*, 125, 389–412.
- Bai, B., Wang, J., & Chai, C. S. (2019). Understanding Hong Kong primary school English teachers' continuance intention to teach with ICT. *Computer Assisted Language Learning*, 1–23.
- Becker, S. A., Cummins, M., Davis, A., Freeman, A., Hall, C. G., & Ananthanarayanan, V. (2017). *NMC horizon report: 2017 higher education edition* (pp. 1–60). The New Media Consortium.
- Cazan, A. M., Cocoradă, E., & Maican, C. I. (2016). Computer anxiety and attitudes towards the computer and the internet with Romanian high-school and university students. *Computers in Human Behavior*, 55, 258–267.
- Chavoshi, A., & Hamidi, H. (2019). Social, individual, technological and pedagogical factors influencing mobile learning acceptance in higher education: A Case from Iran. *Telematics and Informatics*, 38, 133–165.
- Chen, F. H., Looi, C. K., & Chen, W. (2009). Integrating technology in the classroom: a visual conceptualization of teachers' knowledge, goals and beliefs. *Journal of computer assisted learning*, 25(5), 470–488.
- Chen, X., Xie, H., & Hwang, G. J. (2020a). A Multi-perspective study on artificial intelligence in education: Grants, conferences, journals, software tools, institutions, and researchers. *Computers and Education: Artificial Intelligence*, 1, 100005. doi:10.1016/j.caeai.2020.100005
- Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020b). Application and theory gaps during the rise of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 1, 100002. doi:10.1016/j.caeai.2020.100002

- Cheok, M. L., Wong, S. L., Ayub, A. F., & Mahmud, R. (2017). Teachers' perceptions of e-learning in Malaysian secondary schools. *Malaysian Online Journal of Educational Technology*, 5(2), 20–33.
- Clark-Gordon, C. V., Bowman, N. D., Hadden, A. A., & Frisby, B. N. (2019). College instructors and the digital red pen: An Exploratory study of factors influencing the adoption and non-adoption of digital written feedback technologies. *Computers & Education*, 128, 414–426.
- Colchester, K., Hagra, H., Alghazzawi, D., & Aldabbagh, G. (2017). A Survey of artificial intelligence techniques employed for adaptive educational systems within e-learning platforms. *Journal of Artificial Intelligence and Soft Computing Research*, 7(1), 47–64.
- Compeau, D. R., & Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS Quarterly*, 19(2)189–211.
- Dávideková, M., Mjartan, M., & Greguš, M. (2017). Utilization of virtual reality in education of employees in slovakia. *Procedia computer science*, 113, 253–260.
- Davis, F. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340.
- Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A Comparison of two theoretical models. *Management science*, 35(8), 982–1003.
- Edwards, C., Edwards, A., Spence, P. R., & Lin, X. (2018). I, teacher: Using artificial intelligence (AI) and social robots in communication and instruction. *Communication Education*, 67(4), 473–480.
- Farrell, A. M. (2010). Insufficient discriminant validity: A Comment on Bove, Pervan, Beatty, and Shiu (2009). *Journal of Business Research*, 63(3), 324–327.
- Fornell, C., & Larcker, D. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39–50.
- Hsieh, W. M., & Tsai, C. C. (2017). Taiwanese high school teachers' conceptions of mobile learning. *Computers & Education*, 115, 82–95.
- Huang, J., & Chen, Z. (2016). The Research and design of web-based intelligent tutoring system. *International Journal of Multimedia and Ubiquitous Engineering*, 11(6), 337–348.
- Hung, I. C., Yang, X. J., Fang, W. C., Hwang, G. J., & Chen, N. S. (2014). A Context-aware video prompt approach to improving students' in-field reflection levels. *Computers & Education*, 70, 80–91.
- Hwang, G. J. (2003). A Conceptual map model for developing intelligent tutoring systems. *Computers & Education*, 40(3), 217–235.
- Hwang, G. J., Hung, P. H., Chen, N. S., & Liu, G. Z. (2014). Mindtool-assisted in-field learning (MAIL): An Advanced ubiquitous learning project in Taiwan. *Journal of Educational Technology & Society*, 17(2), 4–16.
- Hwang, G. J., Li, K. C., & Lai, C. L. (2020a). Trends and strategies for conducting effective STEM research and applications: A Mobile and ubiquitous learning perspective. *International Journal of Mobile Learning and Organisation*, 14(2), 161–183.
- Hwang, G. J., Sung, H. Y., Chang, S. C., & Huang, X. C. (2020b). A Fuzzy expert system-based adaptive learning approach to improving students' learning performances by considering affective and cognitive factors. *Computers & Education: Artificial Intelligence*, 1, 00003. doi:10.1016/j.caeai.2020.100003
- Hwang, G. J., & Tu, Y. F. (2021). Roles and research trends of artificial intelligence in mathematics education: A Bibliometric mapping analysis and systematic review. *Mathematics*, 9(6), 584. doi:10.3390/math9060584
- Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020c). Vision, challenges, roles and research issues of artificial intelligence in education. *Computers & Education: Artificial Intelligence*, 1, 100001. doi:10.1016/j.caeai.2020.100001
- Hwang, G.-J., Chen, C.-Y., Tsai, P.-S., & Tsai, C.-C. (2011). An Expert system for improving web-based problem-solving ability of students. *Expert Systems with Applications*, 38(7), 8664–8672.
- Ifenthaler, D., & Schweinbenz, V. (2013). The Acceptance of Tablet-PCs in classroom instruction: The Teachers' perspectives. *Computers in human behavior*, 29(3), 525–534.
- Ifinedo, E., Rikala, J., & Hämäläinen, T. (2020). Factors affecting Nigerian teacher educators' technology integration: Considering characteristics, knowledge constructs, ICT practices and beliefs. *Computers & Education*, 146, 103760. doi:10.1016/j.compedu.2019.103760
- Joo, Y. J., Park, S., & Lim, E. (2018). Factors influencing preservice teachers' intention to use technology: TPACK, teacher self-efficacy, and technology acceptance model. *Educational Technology & Society*, 21(3), 48–59.

- Kao, C. P., & Tsai, C. C. (2009). Teachers' attitudes toward web-based professional development, with relation to Internet self-efficacy and beliefs about web-based learning. *Computers & Education*, 53(1), 66–73.
- Kinshuk, Chen, N.-S., Cheng, I.-L., & Chew, S. W. (2016). Evolution is not enough: Revolutionizing current learning environments to smart learning environments. *International Journal of Artificial Intelligence in Education*, 26(2), 561–581.
- Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.
- Lai, H. M., Hsiao, Y. L., & Hsieh, P. J. (2018). The Role of motivation, ability, and opportunity in university teachers' continuance use intention for flipped teaching. *Computers & Education*, 124, 37–50.
- Legrís, P., Ingham, J., & Colletette, P. (2003). Why do people use information technology? A Critical review of the technology acceptance model. *Information & Management*, 40(3), 191–204.
- Lim, C. P., & Khine, M. S. (2006). Managing teachers' barriers to ICT integration in Singapore schools. *Journal of Technology and Teacher Education*, 14(1), 97–125.
- Lin, H. C., Tu, Y. F., Hwang, G. J., & Huang, H. (2021). From precision education to precision medicine. *Educational Technology & Society*, 24(1), 123–137.
- Lin, K. (2011). E-learning continuance intention: Moderating effects of user e-learning experience. *Computers & Education*, 56(6), 515–526.
- Lin, P. H., Wooders, A., Wang, J. T. Y., & Yuan, W. M. (2018). Artificial intelligence, the missing piece of online education? *IEEE Engineering Management Review*, 46(3), 25–28.
- Liu, K. J., Cao, Y. D., Hu, Y., & Wei, L. J. (2019). Application status and development of big data in medical education in China. *Medical Data Mining*, 2(3), 118–125.
- MacCallum, K., & Jeffrey, L. (2014). Comparing the role of ICT literacy and anxiety in the adoption of mobile learning. *Computers in Human Behavior*, 39, 8–19.
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3), 519–530.
- McArthur, D., Lewis, M., & Bishary, M. (2005). The Roles of artificial intelligence in education: current progress and future prospects. *Journal of Educational Technology*, 1(4), 42–80.
- Motaghian, H., Hassanzadeh, A., & Moghadam, D. K. (2013). Factors affecting university instructors' adoption of web-based learning systems: Case study of Iran. *Computers & Education*, 61, 158–167.
- Nelson, M. J., Voithofer, R., & Cheng, S. L. (2019). Mediating factors that influence the technology integration practices of teacher educators. *Computers & Education*, 128, 330–344.
- Nikou, S. A., & Economides, A. A. (2017). Mobile-based assessment: Integrating acceptance and motivational factors into a combined model of self-determination theory and technology acceptance. *Computers in Human Behavior*, 68, 83–95.
- Nilsson, N. J. (2014). *Principles of artificial intelligence*. Burlington, MA: Morgan Kaufmann.
- Oblinger, D. G., ed. 2012. *Game changers: Education and information technology*. Washington, DC: Educause.
- Paechter, M., Maier, B., & Macher, D. (2010). Students' expectations of, and experiences in e-learning: Their relation to learning achievements and course satisfaction. *Computers & education*, 54(1), 222–229.
- Park, C., Kim, D. G., Cho, S., & Han, H. J. (2019). Adoption of multimedia technology for learning and gender difference. *Computers in Human Behavior*, 92, 288–296.
- Popenici, S. A., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, 12(1), 22. doi:10.1186/s41039-017-0062-8
- Rau, M. A., Aleven, V., & Rummel, N. (2017). Making connections among multiple graphical representations of fractions: Sense-making competencies enhance perceptual fluency, but not vice versa. *Instructional Science*, 45(3), 331–357.
- Raykov, T., & Marcoulides, G. A. (2008). *An Introduction to applied multivariate analysis*. New York, NY: Taylor and Francis.
- Russell, S., & Norvig, P. (2003). *Artificial intelligence: A Modern approach (2nd ed.)*. Upper Saddle River, NJ: Pearson Education.
- Sánchez-Prieto, J. C., Olmos-Migueláñez, S., & García-Peñalvo, F. J. (2017). MLearning and pre-service teachers: An Assessment of the behavioral intention using an expanded TAM model. *Computers in Human Behavior*, 72, 644–654.
- Sánchez-Prieto, J. C., Olmos-Migueláñez, S., & García-Peñalvo, F. J. (2016). Informal tools in formal contexts: Development of a model to assess the acceptance of mobile technologies among teachers. *Computers in Human Behavior*, 55, 519–528.

- Scherer, R., Siddiq, F., & Tondeur, J. (2019). The Technology acceptance model (TAM): A Meta-analytic structural equation modeling approach to explaining teachers' adoption of digital technology in education. *Computers & Education*, 128, 13–35.
- Shi, Z. Z., & Zheng, N. N. (2006). Progress and challenge of artificial intelligence. *Journal of computer science and technology*, 21(5), 810–822.
- Teo, T. (2019). Students and teachers' intention to use technology: Assessing their measurement equivalence and structural invariance. *Journal of Educational Computing Research*, 57(1), 201–225.
- Teo, T., & Zhou, M. (2014). Explaining the intention to use technology among university students: A Structural equation modeling approach. *Journal of Computing in Higher education*, 26(2), 124–142.
- Teo, T., Lee, C. B., & Chai, C. S. (2008). Understanding pre-service teachers' computer attitudes: Applying and extending the technology acceptance model. *Journal of computer assisted learning*, 24(2), 128–143.
- Timms, M. J. (2016). Letting artificial intelligence in education out of the box: Educational cobots and smart classrooms. *International Journal of Artificial Intelligence in Education*, 26(2), 701–712.
- Ursavaş, Ö. F., Yalçın, Y., & Bakır, E. (2019). The Effect of subjective norms on preservice and in-service teachers' behavioural intentions to use technology: A Multigroup multimodel study. *British Journal of Educational Technology*, 50(5), 2501–2519.
- van Dinther, M., Dochy, F., Segers, M., & Braeken, J. (2013). The Construct validity and predictive validity of a self-efficacy measure for student teachers in competence-based education. *Studies in Educational Evaluation*, 39(3), 169–179.
- Van Seters, J. R., Ossevoort, M. A., Tramper, J., & Goedhart, M. J. (2012). The Influence of student characteristics on the use of adaptive e-learning material. *Computers & Education*, 58(3), 942–952.
- Wang, W. T., & Wang, C. C. (2009). An Empirical study of instructor adoption of web-based learning systems. *Computers & Education*, 53(3), 761–774.
- Woolf, B. P., Lane, H. C., Chaudhri, V. K., & Kolodner, J. L. (2013). AI grand challenges for education. *AI magazine*, 34(4), 66–84.
- Yeşilyurt, E., Ulaş, A. H., & Akan, D. (2016). Teacher self-efficacy, academic self-efficacy, and computer self-efficacy as predictors of attitude toward applying computer-supported education. *Computers in Human Behavior*, 64, 591–601.
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators? *International Journal of Educational Technology in Higher Education*, 16, 39. doi:10.1186/s41239-019-0171-0

# Prediction of Student Performance in Massive Open Online Courses Using Deep Learning System Based on Learning Behaviors

Chia-An Lee<sup>1</sup>, Jian-Wei Tzeng<sup>2</sup>, Nen-Fu Huang<sup>1</sup> and Yu-Sheng Su<sup>3\*</sup>

<sup>1</sup>Department of Computer Science, National Tsing Hua University, Taiwan // <sup>2</sup>Center for Teaching and Learning Development, National Tsing Hua University, Taiwan // <sup>3</sup>Department of Computer Science and Engineering, National Taiwan Ocean University, Taiwan // ckstar2001@gmail.com // darkdreams0802@gmail.com // nfhuang@cs.nthu.edu.tw // ntouaddisonsu@gmail.com

\*Corresponding author

**ABSTRACT:** Massive open online courses (MOOCs) provide numerous open-access learning resources and allow for self-directed learning. The application of big data and artificial intelligence (AI) in MOOCs help comprehend raw educational data and enrich the learning process for students and instructors. Thus, we created two deep neural network models. The first model predicts learning outcomes on the basis of learning behaviors observed when students watch videos. The second is a novel exercise-based model that predicts if a student will correctly answer examination questions on relevant concepts. The study data were collected from two courses conducted on the National Tsing Hua University's MOOCs platform. The first model accurately evaluated student performance on the basis of their learning behaviors, and the second model efficiently predicted student performance according to how they answered the exercise questions. In conclusion, our AI system remedies the present-day inability of MOOCs to evaluate student performance. Instructors can use the systems to identify poor-performing students and offer them more assistance on a timely basis.

**Keywords:** Learning analytics, Educational big data, Massive open online courses, Artificial intelligence

## 1. Introduction

Massive open online courses (MOOCs), an open-access educational resource available to online learners worldwide, represent a new approach to learning. MOOCs provide not only various study materials and resources but also aid students in self-directed learning. With increasing enrollment in MOOCs, a large amount of learning data has become available for collection and analysis. By harnessing data science, analytics approaches to learning can leverage educational data and help students and instructors enrich their learning processes (Vieira, Parsons, & Byrd, 2018).

Many researchers have analyzed MOOCs data by incorporating big data and artificial intelligence (AI) in their research design. Big data and AI have gained prominence in various fields, including machine learning and data science. Machine learning algorithms are more effective when using larger datasets, and the combination of machine learning and big data has made impressive breakthroughs in data science (Ghahramani, 2015).

In recent times, deep neural networks, an important branch of machine learning, has been used successfully in many AI applications (Su, Chou, Chu, & Yang, 2019; Su, Ni, Li, Lee, & Lin, 2020; Su, Ding, & Chen, 2021). Several researchers have constructed multilayered models that capture more complex features, particularly how online learners learn (Hwang, Sung, Chang, & Huang, 2020; Kastrati, Imran, & Kurti, 2019; Li & Zhou, 2018; Su & Lai, 2021; Yang, Brinton, Joe-Wong, & Chiang, 2017). Boulay (2016), for instance, specified that AI techniques help practitioners better understand learning and pedagogical trends, and related systems help students acquire new skills and grasp new concepts. Therefore, the application of AI to MOOCs has drawn considerable attention in big data analytics. The NMC Horizon Report noted that AI will strengthen the online teaching model, which facilitates adaptive learning and research, and make student–teacher interactions more intuitive and frequent (Adams, Cummins, Davis, Freeman, Hall, & Ananthanarayanan, 2017). Fauvel et al. (2018) designed an AI tool to analyze students' learning effectiveness by collecting learning behavior data with the objective of helping MOOC learners better understand concepts and MOOC instructors deliver more effective courses and offer higher-quality educational tools. AI tools are mainly used to bridge the gap between online learning and physical classes and enable students to achieve their learning goals. Therefore, it is important to personalize MOOC services to students' learning adaptability, habits, and behaviors (Tekin, Braun, & Schaar, 2015).

MOOCs transcend spatial and temporal constraints and have popularized the concept of open education. There is a large quantity of structured and unstructured learning data that are based on learner behavior observations and diverse test questions. The data include personal information (e.g., gender, age, education level, and disciplinary background) and responses to test questions (e.g., number of candidates, number of graduates, number of test

questions, responses to test questions, and evaluation goals). Many scholars have proposed that data analysis can be used to improve a teacher's course and make it more adaptable (Ndukwe & Daniel, 2020; Er, Gomez-Sanchez, Dimitriadis, Bote-Lorenzo, Asensio-Perez, & Alvarez-Alvare, 2019; Lee, 2019; Lu, Huang, Huang, & Yang, 2017; Rupiérrez-Valiente, Munoz-Merino, Diaz, Ruiz, & Kloos, 2017). In MOOCs, learners are free to study the topic of their choice irrespective of time and place. In addition, they do not need to follow the instructor's intended course sequence (Matt, 2018). While the self-regulated learning structure of MOOCs offers considerable flexibility and a wealth of valuable resources, many learners do not complete the courses because of the pressure-free learning environment (Azevedo & Cromley, 2004; Bol & Garner, 2011; Peverly, Brobst, Graham, & Shaw, 2003). MOOCs use self-directed learning as their development model (Li, 2019), and thus, learners must set learning goals and use learning strategies commensurate with their aptitude and background knowledge to master the course content. Through videos, exercises, forums, and other interactive functions, learners must develop appropriate self-regulated learning (Lan, Hou, Qi, & Mattheos, 2019; Matt, 2018). Consequently, to help students achieve classroom success, many researchers have proposed assessment systems that can not only improve students' performance and self-regulation abilities (Lu, Huang, Huang, Lin, Ogata, & Yang, 2018) but also identify scope for improvement in course designs on the basis of students' learning behaviors.

There is growing literature on the application of big data in education. Processing large quantities of learning data can elucidate the relationship between learning behaviors and learning effectiveness, which can help educators forecast learning outcomes (Hwang, Chu, & Yin, 2017). The conceptual framework underlying learning analytics can be used to analyze course characteristics, assess student performance, and predict learning progress. According to Lu et al. (2018), learning analytics help educators save time, which can be used to refine their teaching expertise and identify at-risk students at an earlier stage. However, MOOCs have fundamental problems. The most well-known being the low completion rate and the lack of learning guidelines (Freitas, Morgan, & Gibson, 2015). There are varying factors attributable to low completion rates. However, studies have reported that most MOOC learners are unprepared for the extensive course content and isolated learning environment (Kim, Olfman, Ryan, & Eryilmaz, 2014).

In order to address these issues, this study aimed to develop an AI-based system that helps teachers better understand their students' learning performance. The system has two functions. First, it analyzes students' learning behaviors to evaluate their learning performance at a given time. Second, it uses a novel exercise-based model to predict if students will correctly answer examination questions on relevant concepts. We first collected data on the video-watching behavior of participating MOOC students and the frequency at which students watched the videos. These data were subsequently analyzed and used to predict students' scores. The scores calculated using our formulated neural networks can be used to identify students with learning difficulties, the key practical implication of this feature. Previous studies have indicated the following challenges in developing intelligent tutoring systems: techniques that simulate the intelligence of human experts and the need for human tutors' knowledge and experience to make judgments and decisions using the best available evidence to help solve learners' problems and improve their learning ability (Hwang, Xie, Wah, & Gašević, 2020). Second, we collected students' answers to exercises and data on their answering process. Using the data, our system predicted whether a student would answer an examination question correctly.

The system is based on deep learning, a promising technology applied in the field of education. While there has been growing interest in AI-based education research since 2001, less than 5% of such studies focus on AIED. However, considering its rapid advancement, there is much potential in the application of deep learning in education (Chen, Xie, & Hwang, 2020). Therefore, our proposed system could exemplify the development of a deep learning system to predict student performance.

Finally, most software tools based on AI technologies used for educational purposes are designed to learn languages or mathematics (Chen, Xie, Zou, & Hwang, 2020). The data used for this study are collected from two MOOCs courses: *Introduction to IoT* (where IoT refers to the Internet of Things) and *Calculus I*. Both are introductory courses. The former is for computer science undergraduates from the National Tsing Hua University (NTHU) and covers related techniques. Therefore, in light of future research, the system proposed in this study can be used for programming learning purposes, an arguably important advance in artificial intelligence in education research.

The present AI-based system used NTHU's MOOCs platform as an experimental site. Its objective is to provide teachers with accurate evaluations to identify students with learning difficulties. Furthermore, the predicted results for students' examination answers could help teachers understand students' learning experience without the need to conduct additional exams. Consequently, teachers may be able to better guide their students and increase their motivation to learn. This study was based on the following research questions:



**RQ1.** In a MOOC learning environment, can video-watching data that reflect learning behaviors be used to evaluate learning outcomes in addition to online assessment scores (e.g., quiz or examination scores)?

**RQ2.** In addition to the proportion of correctly answered questions, can deep learning be applied to the aforementioned video-watching data to evaluate if a student has mastered the course content and understood related concepts?

## **2. Literature review**

### **2.1 Data analysis and enhancement of learning effectiveness**

AI refers to the simulation of human intelligence in machines such that their judgments and decisions exhibit the characteristics of a human mind (Akerkar, 2014; Su, Ding, & Chen, 2021; Su, Suen, & Hung, 2021). In recent years, research on artificial intelligence in education (AIED) has flourished with the increasing sophistication of data analytics (Kay & Kummerfeld, 2019; Schwendimann, 2017; Su & Lai, 2021; Su & Wu, 2021). The literature has also witnessed the development of new research methods and subfields, such as educational data mining and learning analytics, where scholars gather learner data from online platforms to analyze learning processes (Daghestani, Ibrahim, Al-Towirgi, & Salman, 2020; Alexandron, Ruipérez-Valiente, Chen, Muñoz-Merino, & Pritchard, 2017; Romero & Ventura, 2017).

The proliferation of data analytics, especially big data analysis, in education has paved the way for a new teaching approach, wherein student activities and progress are tracked to improve learning outcomes. In addition, students can track their learning progress for better self-directed learning (Alonso-Mencia et al., 2019; Kavitha & Raj, 2017). Hwang et al. (2020) developed a fuzzy expert system-based adaptive learning approach while accounting for both affective and cognitive factors. The experiment results indicated that the learning system could enhance students' learning achievements and reduce their learning anxiety.

Advances in learning data analytics have led to the creation of an accommodating online learning environment that helps students achieve their learning goals, especially in higher education distance teaching and teacher training courses. Using such technologies, teachers can track learning behaviors and evaluate students' learning effectiveness across several dimensions (Meier, Xu, Atan, & Schaar, 2016).

### **2.2. Evaluation of learning performance based on student behavior**

Learning behaviors are learned actions commonly used to assess students' learning and performance. Examining students' learning behaviors not only gives teachers insight into students' learning situations, but also ensures the feasibility of teaching materials. Hsu et al. (2021) developed an instructional tool for AI education and used videos and screenshots to record learning behaviors. Their study revealed meaningful behavioral patterns when students learned the application of AI.

Students' learning behaviors on MOOCs are also an important factor in learning assessments. MOOCs, however, commonly report low completion and high dropout rates (Sun, Ni, Zhao, Shen, & Wang, 2019). Numerous studies have proposed methods to predict students' success or failure in courses (Er et al., 2019; Lu et al., 2018). One such method uses a logistic regression model for prediction. Lee (2018) applied this method to analyze the behavior of students engaged in uninterrupted video watching and examined data drawn from the students' learning logs. Students reported interrupted learning if they did not watch the course video for two consecutive days. The author estimated the frequency and duration of uninterrupted learning actions from the learning log data and inputted the data into the prediction model. Lee then defined three thresholds for continual learning (10, 30, and 60 minutes) and compared the effect of uninterrupted learning across the thresholds. The 60-minute threshold occupied the largest area under the precision-recall curve, indicating that the threshold was the most useful in predicting student success in obtaining a course certificate. In other words, students are more likely to obtain a course certificate if they participated in more learning activities and engaged in learning for a longer duration.

Guo, Kim, and Rubin (2014) proposed several features that educational video production should incorporate to increase engagement, which was measured by the duration of video watching and whether students attempted a post-video exercise. Using simple statistical tools, the authors found that shorter videos, in addition to other video production decisions, led to greater engagement. These findings can be useful for MOOC instructors.

Kim et al. (2014) revealed that video length was strongly and negatively correlated with engagement; that is, learners were less likely to finish watching a longer video. The authors also demonstrated that students were more likely to view the entire video when they watched it for the first time rather than when they did so more than once. Using binning and kernel-based smoothing, the authors then produced second-by-second plots of peaks in video interactions (defined by play, pause, and skip). The plots revealed students' learning behaviors when they watched a video. Each peak was manually classified into five categories to explain the underlying cause of the peak. Their results elucidated how students interact and learn, and practitioners can use these findings to improve video interfaces for learning.

Sun et al. (2019) proposed a gated recurrent unit-recurrent neural network (GRU-RNN) model to construct a dropout prediction model. The model is based on an RNN with a URL embedding layer. The authors used their model to compare student performance before and after course entry and to determine the number of days students did not spend on learning. They then analyzed different approaches to learning, such as answering exercise questions, interacting on forums, and taking examinations. Finally, the authors examined students' learning habits through their sequence of learning behaviors to predict learning performance.

### **2.3. Measurement of learner proficiency in MOOCs**

Traditional learning assessments offer a judgment score or standard reference. However, students differ in their learning ability and speed. Difficult test questions poorly reflect the comprehension level of students with low learning ability. To address this issue, researchers formulated test response theory, which became increasingly popular in education research and practice. According to test response theory, students receive questions on the basis of their response to the previous ones, and thus, the difficulty level of the test is tailored to a student's ability. However, the theory does not address ways to dispel student misconceptions or to diagnose learning disabilities (Liu, Lin, & Tsai, 2009). There are several methods to conduct a diagnosis. Interviews are the most common qualitative method, and test response theory is the most frequently used quantitative method. With the growing application of AI technology, including neural networks, diagnostic testing is an emerging subfield in the testing industry. Chu (2020) envisioned cognitive diagnostic testing that is based on cognitive science theory as a crucial future trend. The author designed a cognitive diagnostic test and proposed a question-response model to verify if cognitive science theory yields valid evaluations for student ability (Chu, Li, & Yu, 2020). Their method helped improve learning data analytics, thus allowing MOOC teachers to better evaluate student performance and track learning behaviors across various learning dimensions.

The MOOC literature has widely investigated online assessments and learner participation. DeBoer, Ho, Stump, and Breslow (2014) analyzed the concept of participation and desirable metrics for learning success and participation quality. However, learners might sign up for a course and not complete the assessments. Admiraal, Huisman, and Van de Ven (2014) explored the assessment quality of MOOCs. MOOCs entail a dynamic learning process: learners engage in a series of actions comprising perception, learning, thinking, and problem-solving. Thus, final scores are an inadequate indicator of learner performance (Shepard, 2001). Teachers must observe students' learning behaviors during a course since learning is a process rather than an outcome. The aforementioned conclusions emphasize the need for alternative assessment methods in MOOCs.

### **2.4. Prediction of learning performance using exercises**

Moreno-Marcos, Pong, Muñoz-Merino, and Delgado-Kloos (2020) presented a method to predict students' assignment, examination, and final grades on the basis of their learning status, performance in discussion forums, video-watching behaviors, answers to practice questions, and previous assignment scores. The authors found that previous assignment scores and average answer scores were highly predictive of the aforementioned three grades, whereas student performance in discussion forums was only slightly predictive. Because some courses provide videos without exercises, the authors used student behavioral data such as click counts as a model feature but noted no substantial change in performance.

Learning styles in MOOCs can be categorized by performance in course assessments (Alario-Hoyos, Pérez-Sanagustín, Delgado-Kloos, Parada, & Muñoz-Organero, 2014). Alario-Hoyos et al. (2014) used learner performance in a sequence of course activities (including videos and exercises) to cluster learners into three broad categories: lurkers, participants who did not complete a course, and participants who completed the course. Although the authors did not detail their clustering method, it appeared to be based on simple statistics.

Ashenafi, Riccardi, and Ronchetti (2015) proposed a method to predict the final examination results of students in two undergraduate programming courses (*Informatica Generale I* (IG1) and *Programmazione II* (PR2)) at the University of Trento, Italy. Throughout the courses, students participated in a set of peer-based online homework activities with three main tasks: ask a question, answer a question, and rate answers. A total of 14 types of data were captured before they were used as input features in a prediction model with logistic regression. The prediction model outperformed its counterparts by a root mean square error of 2.93 for one course and 3.44 for the other.

Huang, Chen, Tzeng, and Lee (2018) designed a concept assessment system with a knowledge map using deep learning. The authors presented each week's knowledge topology as a knowledge map. They collected data on the difficulty level of exercises and student behaviors when watching videos and used the data to predict students' comprehension of the content in a given week's course. The prediction model was based on a deep learning method.

Li, Xie, and Wang (2016) proposed a model to predict test scores. Drawing on several educational theories, the authors predicted quiz grades using 15 features such as student age, gender, education level, registration time, number of videos watched, number of exercises, and related actions. However, the features were not significantly associated with examination scores, and thus, could not be used in the prediction model.

## 2.5. Lack of evaluation mechanisms in MOOCs

Student performance has been traditionally evaluated using standardized tests, and thus, there is a need for learning tools that evaluate learning investments in hybrid, remote, or virtual learning environments. MOOCs have altered global learning trends, although they face many challenges in terms of their long-term development and learning models, including low completion rates (5-10% on average) and high learning loss rates (Sun, Ni, Zhao, Shen, & Wang, 2019). Evaluating learner performance in MOOCs is inherently difficult because students cannot be monitored in real-time, limiting MOOCs' ability to be impartial or provide reliable proof of coursework (Bady, 2013). Moreover, MOOCs have numerous learners, and teachers cannot interact with every student. In such cases, students must rely on active interactions with other online learners to obtain learning feedback and practice. Importantly, students must be self-directed learners (Crosslin, 2018). Previous evaluation methods for online learners can serve as a guide for educators; however, MOOC educators are seeking to develop online metrics for large-scale data collection for students of different levels and ages. Table 1 summarizes missing components in MOOC assessments, factors contributing to these gaps, and how these gaps can be bridged with our deep learning system.

Table 1. Lack of assessment in MOOCs: Reasons and proposed solutions

Learning problem	Reason	Solution
Assessment is potentially unfair.	Students cannot be monitored in real-time, and there is scope to cheat on tests.	Our system performs a big data analysis to provide MOOC educators with an evaluation system that supplements examinations.
Examinations do not provide clear and objective evaluations.	MOOC learners are diverse, and some may have inadequate background knowledge for a course.	Our system uses neural networks to estimate objective and credible evaluation scores using large datasets on learning behaviors and judgments.
Effective learning feedback is lacking.	Different learners absorb different content.	Our system draws on learning behaviors to predict the proportion of questions students will answer correctly. These predictions will help teachers understand if students have grasped related concepts.

## 3. Methods

This section describes the use of data on video-watching behaviors and answers in exercises to predict students' learning performance in MOOCs.

### 3.1. Course information and collection of data on learning behaviors

Students from two MOOCs courses participated in this study. Table 2 details the two courses. Students must obtain a minimum score of 60 to complete either course.

The introductory course for IoT is for computer science undergraduates at NTHU and covers techniques used in IoT. Students are expected to spend three hours per week watching online videos and to participate in offline laboratory sessions during which they can conduct experiments. Students can complete exercises as practice and discuss the course content with their peers on the online platform.

The 12-week comprehensive introduction to calculus is a prerequisite for all first-year students and must be completed during the summer vacation. Students are expected to spend three hours per week watching videos and to complete relevant exercises.

Table 2. Course information

	Introduction to IoT	Calculus I
Duration	March 2–June 29, 2020	May 1–August 31, 2020
Number of students	255	1,062
Number of videos	87	144
Number of weeks	5	12
Average video time	525	792
Number of exercises	71	143
Number of quizzes	1	3
Number of questions per quiz	50	20
Quizzes interval time (in weeks)	5	4
Course qualification	No	High school students only
Fee	Free	Paid

Videos constitute the primary teaching method in most MOOCs. For this study, we collected data on video playback actions, such as play, pause, search, and adjust playback speed, on the YouTube application programming interface (API) and then stored the data on the MongoDB database (Table 3). In addition, we collected data on each user's answers for all exercises (Table 4). If students navigated to the exercise page but did not answer the exercise questions, we coded student responses to the exercise as "no." The "timeCost" feature is the duration students took to answer a question. For example, if a student spent 20 seconds answering a question, the timeCost value for the question was 20.

Table 3. Student video activity schema

	Description	Example
userId	Student ID	2,556
courseId	Course ID	10900MATH0001
chapterId	Chapter ID	10900MATH0001ch79
videoId	Video ID	-RHQ75vrT3Q
Action	Student action when recording	Playing
currentTime	Video time when recording	29.57483
playRate	Video play rate when recording	1.25
Volume	Video volume when recording	100
update_at	Recording time	2020-05-20T15:48:03

Table 4. Student exercise activity schema

	Description	Example
userId	Student ID	2,556
courseId	Course ID	10900MATH0001
chapterId	Chapter ID	10900MATH0001ch79
exerId	Exercise ID	10900MATH0001ch79e1
score	Exercise answer score	0.6
timeCost	Time cost on exercise	15
userAns	User answer	[1, 3]
correctAns	Correct answer	[1, 2, 3]
update_at	Recording time	2020-05-15T09:33:35

### 3.2. Learning variables: Video-watching frequency and duration

This subsection presents the definition of the variables used in this study: frequency and duration of video watching (Table 5). The variables are associated with a given day: on such a day, students primarily learned by watching videos. The average duration of a video is 10–15 minutes. We considered students to have engaged in learning if they watched a video for more than 5 minutes. Figure 1 is an example of a student’s video-watching log.

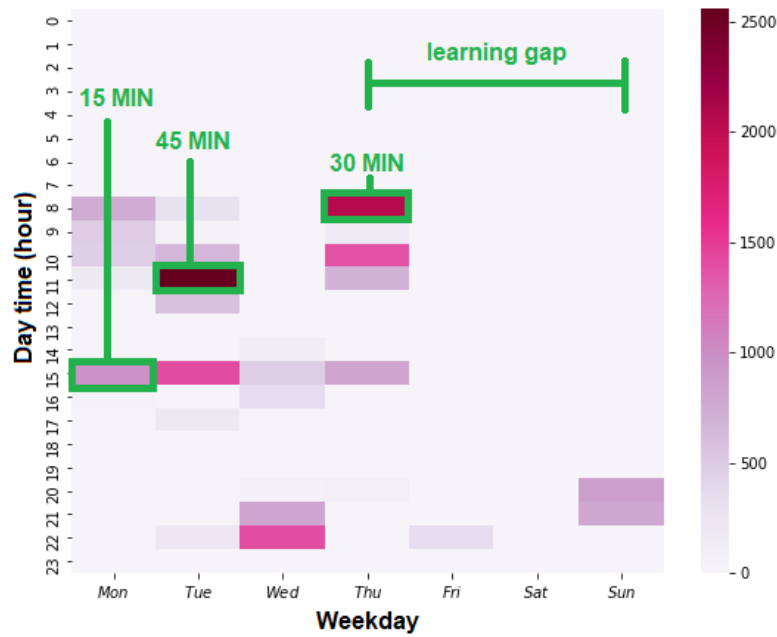


Figure 1. Student video-watching log

Table 5. Video-watching features

Features	Description
<i>videoFinishRate</i>	Proportion of videos finished
<i>videoSpendTime</i>	Time spent watching videos/total time of all videos
<i>Play</i>	Mean of playing in watching videos per week
<i>gapMean</i>	Mean of days not spent on learning per week
<i>gapStd</i>	Standard deviation of <i>gapMean</i>
<i>regDay</i>	Number of days per week spent on learning
<i>weekBlockNumMean</i>	Mean number of learning blocks per week
<i>weekBlockNumStd</i>	Standard deviation of <i>weekBlockNumMean</i>
<i>weekBlockTimeMean</i>	Mean time of learning blocks per week
<i>dayBlockNumMean</i>	Mean number of learning blocks per learning day
<i>dayBlockNumStd</i>	Standard deviation of <i>dayBlockNumMean</i>
<i>dayBlockTimeMean</i>	Mean time of learning blocks per learning day
<i>15Min</i>	Mean number of learning blocks >15 minutes/week
<i>30Min</i>	Mean number of learning blocks >30 minutes/week
<i>45Min</i>	Mean number of learning blocks >45 minutes/week
<i>weekNum</i>	Weeks since course started

#### 3.2.1. Video-watching behavior

In addition, we defined variables for each student that captured their behaviors when watching a video. The variables were the proportion of course videos a student finished watching and that of total video playback time. The second variable was calculated as  $1 - (a / b)$ , where  $a$  is the total playback time for parts of all videos a student did not watch and  $b$  is total playback time for all course videos.

### 3.2.2. Learning gap

A learning gap refers to the number of days a student did not spend on learning and was used to indicate a student's learning pace.

### 3.2.3. Uninterrupted learning

A learning block constitutes uninterrupted periods of learning. We estimated the number of learning blocks for each student and the duration of the learning blocks per day or week. We set three time thresholds as per the length of the videos.

### 3.2.4. Learning regularity

We determined whether a student was learning regularly. To denote such regularity, we first recorded if a student had a dedicated learning day per week throughout the semester. We then aggregated the total number of such days. However, we also found some students dedicating learning days closer to the examination rather than throughout the semester. In other words, they “crammed” their learning, and such students were given the lowest regularity value (−1).

## 3.3. Learning variable: Answers to exercise questions

We recorded and analyzed each student's answer to all exercise questions and extracted eight features (Table 6).

Table 6. Exercise features

Features	Description	Example
Exercise type	Exercise type (single, multiple, fill in the blanks)	Multiple
Correct rate	Percentage of questions answered correctly	0.1
Answer count	Number of attempts before student answers correctly	3
Time cost	Time taken to complete exercise	60
Pre-answer review	Whether student watched a related video before answering correctly the first time	False
Post-answer review	Whether student watched a related video after answering correctly the first time	True
Answering process	Type of question-processing style (type 1–6)	5
Correct count	Number of questions answered correctly	0

### 3.3.1. Rate of correctly answered questions

The rate of correctly answered questions indicated the difficulty level of an exercise. We use this indicator because the difficulty levels of exercises are not always defined by the test creator, and not all students have similar learning abilities. A higher number of correctly answered questions denotes greater student proficiency.

### 3.3.2. Number of attempts before correctly answering a question the first time

The number of attempts before correctly answering a question for the first time indicates the difficulty level of an exercise, where a greater number indicates a higher difficulty level. However, this feature may be directly affected by the difficulty level of an exercise.

### 3.3.3. Watching related videos before or after correctly answering the first time

If students watched videos related to a question within 10 minutes of answering correctly the first time, we defined them as having an impression of relevant concepts when attempting the exercise. By contrast, if students watched related videos within 10 minutes of finishing the exercise, we defined them as being unfamiliar with the concepts and indicated that they gained familiarity only after watching the videos.

### 3.3.4. Student approach to questions before answering correctly the first time

We collected data on student behaviors before correctly answering a question the first time. Students were divided into six types depending on how they processed the answers (Table 7).

Table 7. Types of students based on answering process

Answering Process	Attempt Count Before Answering Correctly First Time	Incorrect Answer Count Before Answering Correctly First Time	Final Result (Correct or Incorrect)	Example
1	1	0	True	[C]
2	2	1	True	[W, C]
3	>2	>1	True	[W, W, C]
4	>1	0	True	[no, no, C]
5	>0	>0	False	[W, no, W]
6	≥0	0	False	[no, no, no]

Note. C = correct answer, W = wrong answer, no = skipped question.

### 3.3.5. Number of correct answers

Except for the number of correctly answered questions, all the aforementioned features are related to student behaviors when correctly answering a question for the first time. These represent a student's proficiency in corresponding knowledge nodes, as formulated by Muñoz-Merino, Ruipérez-Valiente, Alario-Hoyos, Pérez-Sanagustín, and Kloos (2015), who also mentioned that the repeated practice of exercise questions improves student learning and achievement. While exercises on NTHU's MOOC platform are not in parametric form (and thus, such repeated practice is less effective), we believe the number of correct answers represents a student's perception of how much information an exercise contains.

## 3.4. Prediction of learning performance based on video-watching behaviors

Every student has a unique learning mode and behavior, and we hypothesized that these affect their learning performance. To verify this hypothesis, we fed data on learning blocks, gaps, and regularity into a deep neural network (DNN) model. The model used ReLU as the activation function to predict student performance. Note that when creating predictions in MOOCs, it is necessary to avoid inaccuracies caused by sparse data (Yang et al., 2017). To resolve this problem, we only incorporated learning data for students who took the quiz in our system. Figure 2 illustrates the architecture of our performance prediction model, including the features we used and the number of nodes in each DNN layer. The mean absolute error (MAE) was applied to denote the model's performance. In brief, we used 10-fold cross-validation and shuffling to obtain test data. The data were then used to calculate the MAE as follows:

$$MAE = \frac{1}{N} \sum_{i=1}^N |(f_i - y_i)|,$$

where  $f_i$  and  $y_i$  are the predicted and actual scores of student  $i$ , and  $N$  is the number of students. MAE is the difference between the predicted and actual scores, with a lower MAE indicating better predictive performance. The number of hidden layers was determined using trial and error and cross-validation in performance tests for DNN (Table 8). Figure 3 shows training and validation loss during the training of the predication model on the basis of video-watching behaviors.

Table 8. Number of hidden layers vs. mean absolute error

	6 layers	7 layers	8 layers
MAE	8.5	7.7	6.8

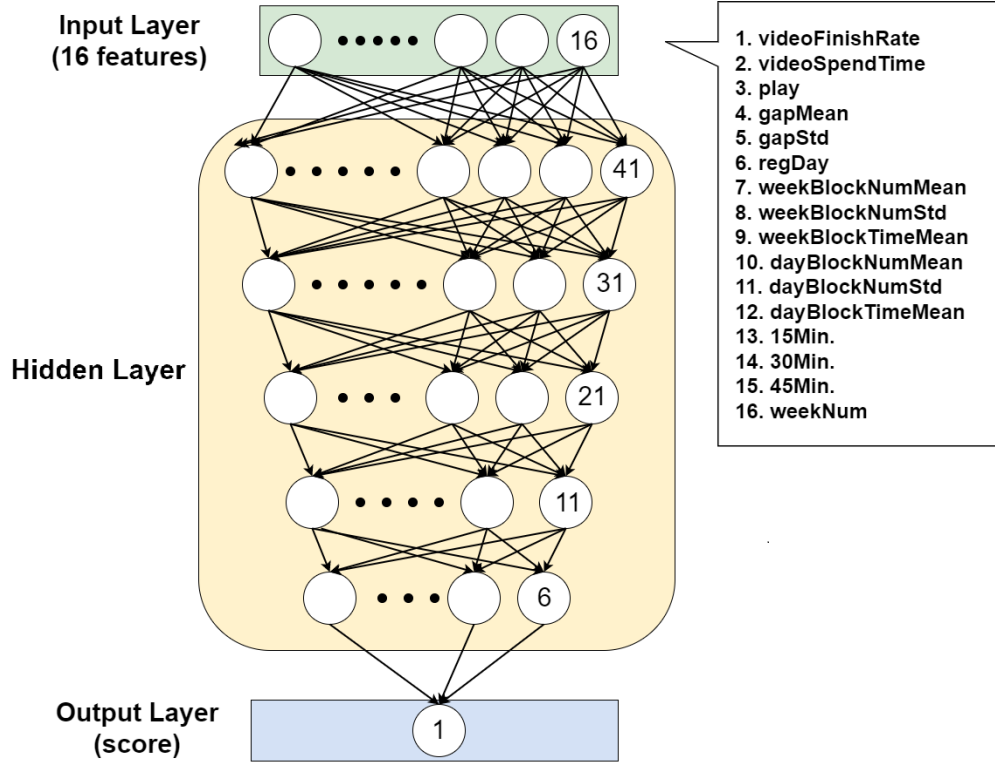


Figure 2. Architecture of score prediction model based on video-watching behaviors

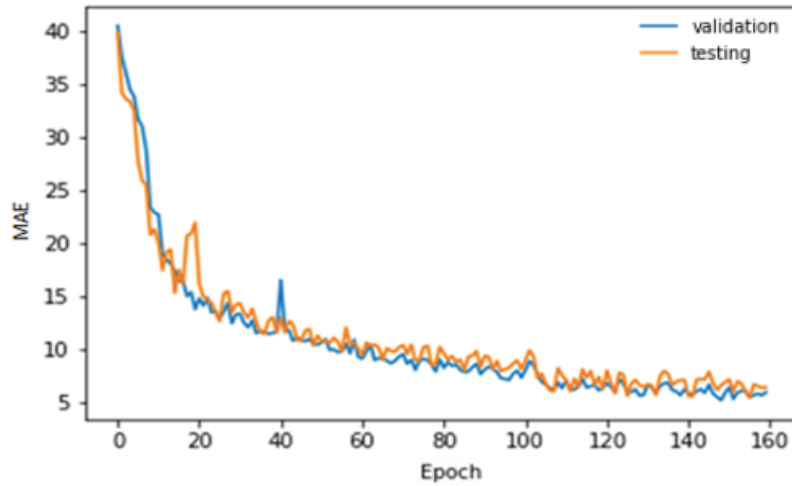


Figure 3. Learning curve for prediction model based on video-watching behaviors

### 3.5. Prediction of familiarity with concepts based on answers to exercises

Exercises potentially indicate if a student is familiar with a course's content. Thus, in this study, we input the aforementioned variables on students' exercise-answering behavior in a five-layered DNN model (Figure 4) to predict learning performance. Table 9 shows the number of hidden layers determined using trial and error and cross-validation. We defined a large number of such variables (e.g., number of attempts, videos watched before answering correctly, and rate of correctly answered questions) to obtain better predictions. We then used the sigmoid function as the activation function to determine the probability of a correct answer. We set the threshold to 0.5, and if the probability of a correctly answered question is greater than or equal to 0.5, then the answer can be judged as correct (and vice versa). Figure 5 shows training and validation loss during the training of the prediction model on the basis of exercise-answering behaviors.



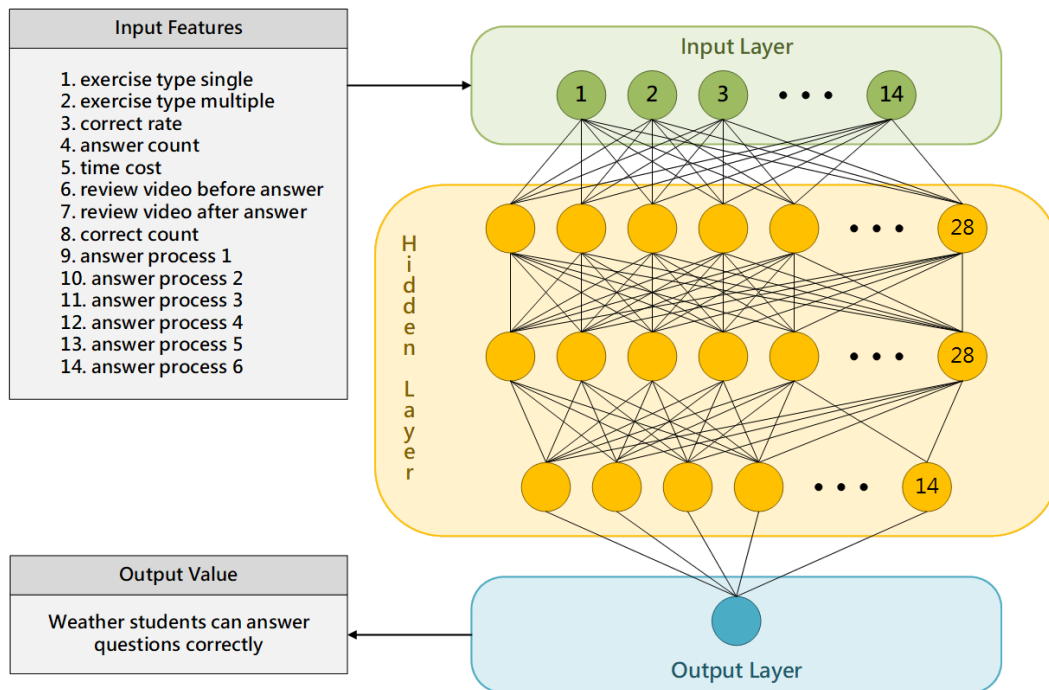


Figure 4. Architecture of prediction model based on exercise-answering behaviors

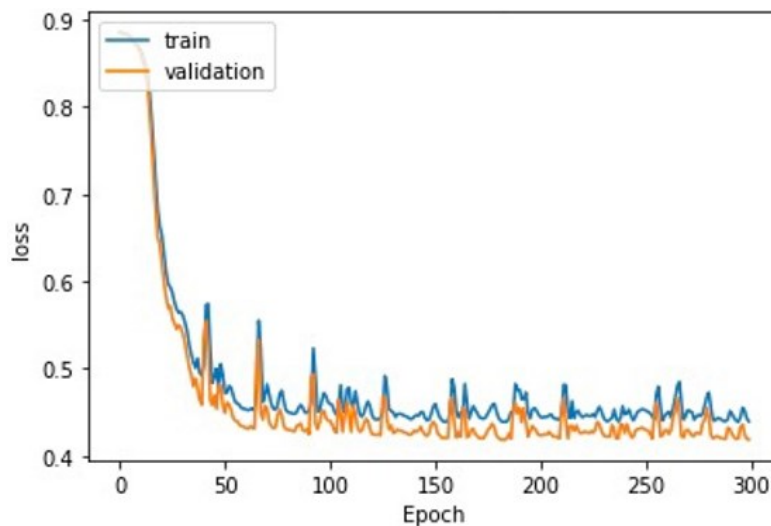


Figure 5. Learning curve of prediction model based on exercise-answering behaviors

Table 9. Number of hidden layers vs. accuracy

	4 layers	5 layers	6 layers
Accuracy	0.733	0.975	0.883

## 4. Results

### 4.1. Use of learning behaviors to evaluate learning performance

The introductory course on IoT conducts a final exam to evaluate student performance, whereas the calculus course administers a quiz every four weeks (a total of three quizzes). We built temporal performance prediction models on a weekly basis to measure the accuracy of the prediction approaches. We ran our models at the end of each week. “Week1” represents all collected log data from the beginning of the course to the end of the first week, and “Week2” denotes data collected from the start of the course to the end of the second. We construct similar variables for the remaining weeks of the two courses. We then verified the effectiveness of our prediction model on the basis of MAE using data for the two courses. Since the introductory IoT course conducted one final exam to evaluate students’ performance, every predicted result is validated by the same actual data (students’

final exam score). Therefore, Table 10 contains only one MAE. On the other hand, the calculus course had three tests, and thus, Table 11 comprises three MAEs. “Quiz1-MAE” is the comparison between our predicted scores and the actual scores of the first quiz. “Quiz2-MAE” is a comparison of our predicted scores with the actual scores of the second quiz, and “Quiz3-MAE” is a comparison of our predicted scores and actual scores of the third quiz.

*Table 10. Predictive performance (MAE) for IoT introduction course*

	MAE
Week1	18.2
Week2	13.6
Week3	10.1
Week4	7.9
Week5	6.9

*Table 11. Predictive performance (MAE) for calculus-I course*

	Quiz1-MAE	Quiz2-MAE	Quiz3-MAE
Week1	17.9	22.93	23.7
Week2	15.6	22.3	23.1
Week3	10.56	21.4	22.6
Week4	6.9	19.3	21.7
Week5	X	15.9	17.4
Week6	X	12.0	12.2
Week7	X	8.4	8.9
Week8	X	7.0	8.0
Week9	X	X	7.19
Week10	X	X	6.9
Week11	X	X	6.8
Week12	X	X	6.8

Table 10 shows a significant gap between students’ predicted scores at the beginning of the IoT introductory course and the actual final scores. However, our model’s performance improved in the following weeks. MAE based on Week5 (students’ whole learning behavior) was only 6.9 points, indicating that our model had acceptable accuracy.

Table 11 shows that Quiz1-MAE based on Week4, Quiz2-MAE based on Week8, and Quiz3-MAE based on Week12 are all less than seven points, indicating that the completeness of the collected data affected our model’s accuracy. That is, for a given test administered in WeekN of a course, our model’s prediction would have the least errors if its input was WeekN.

For **RQ1**, since all the above-mentioned MAEs are less than seven points, it is reasonable to conclude that our model can accurately predict student performance in a given course on the basis of their learning behavior. Accordingly, running our model on a weekly basis could give teachers reliable information on student performance at the end of each week.

The conclusion also supports that our system is an alternative approach that teachers can adopt to track student performance without repeatedly administering tests. Teachers can use the model to identify students who may need more teaching assistance and accordingly, provide such aid on a timely basis. Finally, this model could enable students who have failed courses to identify changes they need to make to their learning patterns.

#### 4.2. Use of exercise data to predict learning performance

Using the features mentioned in Table 6, the exercise-based model could predict students’ familiarity with concepts when answering exercise questions. In other words, this model could predict if a student would correctly answer a question on relevant concepts by collecting and analyzing students’ answer records. Table 12 lists the number of times two Calculus I students (students C and D) answered the quiz questions correctly and incorrectly, along with the predicted result. Finally, we applied a confusion matrix (Table 13) to the model to estimate the model’s accuracy, recall, precision, and F1 score. All the aforementioned values were acceptable, indicating that the exercise-based model had acceptable predictive power.

Table 12. Comparison of predicted and actual results for two calculus-I students

Student C			Student D	
	Real	Predict	Real	Predict
Question 1	Correct	Correct	Correct	Correct
Question 2	Correct	Correct	Correct	Correct
Question 3	Correct	Correct	Correct	Wrong
Question 4	Wrong	Wrong	Correct	Correct
Question 5	Correct	Correct	Wrong	Wrong
Question 6	Wrong	Wrong	Wrong	Wrong
Question 7	Correct	Correct	Wrong	Wrong

Table 13. Confusion matrix of predicted results for calculus-I students by exercise-based model

	Predicted wrong	Predicted correct
Actual Wrong	6,386	0
Actual Correct	173	418
Accuracy		0.975
Precision		1
Recall		0.707
F1-Score		0.828

Regarding **RQ2**, the confusion matrix results indicated that our system with the exercise-answering feature could provide high-quality predictions. MOOC instructors who use online exercises can feed answering data into our system to better understand how students learn.

### 4.3. Comparisons with other models

#### 4.3.1 Use of learning behaviors to evaluate learning performance

We compared our research with another model by building a baseline model and using the same data as input. We referenced Python’s scikit-learn (sklearn) library to build the SVR baseline model, and set the kernel parameter as “rbf.” Table 14 lists the most critical MAEs in this baseline model.

Table 14. Support vector regression predictive performance

Calculus I	MAE
Week4: Quiz 1	15.6
Week8: Quiz 2	15.2
Week12: Quiz 3	15.6
IoT Introduction	MAE
Week5: Quiz	20.3

#### 4.3.2. Use of exercise data to predict learning performance

Similarly, we deployed a decision tree model with the same data as input to predict if the students would correctly answer questions using relevant knowledge. We referenced Python’s sklearn library to build the decision tree baseline model. We set the criterion parameter to “gini.” Table 15 presents the predicted results for the decision tree model.

Table 15. Predicted results for calculus-I students by decision tree model

Accuracy	0.812
Precision	0.801
Recall	0.603
F1-Score	0.695

## 5. Conclusions

In this study, we designed a system with two functions to help teachers better understand students’ learning performance. The first function evaluated student performance on the basis of their learning behaviors. We tested

our system using student data from two courses conducted on NTHU's MOOCs platform. The data included students' video-watching behaviors and answering exercise questions. We formulated a deep learning model, which processed the data and estimated a predicted grade for each student. The study indicated that the model needed a complete overview of students' learning behavior to obtain the most accurate outcome. The second function used an exercise-based DNN model to effectively evaluate a student's performance on the basis of how they answered exercise questions.

(1) Recent research highlighted the problems of high dropout and low completion rates for MOOCs (Sun, Ni, Zhao, Shen, & Wang, 2019). Since MOOCs are a public online course platform, some students may cheat on an exam, and thus, it is difficult to ensure that students consistently follow the honor code. Therefore, MOOCs may not be a fair learning environment. Moreover, questions have been raised about the authenticity of course credits and certificates (Bady, 2013).

(2) Therefore, this study aimed to propose an objective and accurate AI-based method to examine students' learning effectiveness without interference in MOOCs.

(3) In addition, the proposed model could give teachers more accurate information on whether students have mastered a concept. Our system used the scores for video-watching behaviors and accuracy scores for assigned quizzes and final exams to reflect students' learning outcomes.

In conclusion, our AI system could remedy the present-day inability of MOOCs to evaluate student performance on the basis of learning behaviors, which is a major contribution of our study, particularly to the creation of precision education platforms. Importantly, the experimental results of our model were significantly better than those of the baseline models. The results sufficiently demonstrated the feasibility of using DNN. Instructors can use our systems to identify low-performing students and provide them with additional support. By doing so, our system may create a learning-teaching environment that benefits both students and lecturers.

Despite the valuable findings, our study is subject to certain limitations because of the constraints in time and testing frequency. We focused on two MOOC courses, and these courses did not administer quizzes every week. In addition, we only used student behaviors to evaluate their performance.

Therefore, future studies should consider applying the proposed AI-based evaluation system to other MOOCs to validate its effectiveness using larger datasets. The improved system could incorporate the feature of sending notifications to students to help them accurately evaluate their current study patterns before a course ends. This would give them the opportunity to optimize their learning behaviors. Finally, future works could combine other affect-detecting systems such as student response systems (Li, & Wong, 2020) with our proposed system to obtain real-time affective factors. By analyzing student responses, teachers can take prompt action to improve learning and teaching (Hwang et al., 2020).

## Acknowledgement

This study was funded by the Ministry of Science and Technology of Taiwan. The projects ID were MOST-109-2511-H-007-010, MOST 109-2511-H-019-004-MY2, MOST 109-2511-H-019-001, and MOST-108-2221-E-007-062-MY3. It was also funded by National Tsing Hua University, Taiwan.

## References

- Adams Becker, S., Cummins, M., Davis, A., Freeman, A., Hall Giesinger, C., & Ananthanarayanan, V. (2017). *NMC Horizon Report: 2017 Higher Education Edition*. Austin, TX: The New Media Consortium.
- Admiraal, W., Huisman, B., & Van de Ven, M. (2014). Self-and peer assessment in massive open online courses. *International Journal of Higher Education*, 3(3), 119-128. doi:10.5430/ijhe.v3n3p119
- Akerkar, R. (2014). *Introduction to artificial intelligence*. Patparganj, India: PHI Learning Pvt. Ltd.
- Alario-Hoyos, C., Pérez-Sanagustín, M., Delgado-Kloos, C., Parada G, H. A., & Muñoz-Organero, M. (2014). Delving into participants' profiles and use of social tools in MOOCs. *IEEE Transactions on Learning Technologies*, 7(3), 260-266. doi:10.1109/TLT.2014.2311807
- Alexandron, G., Ruiperez-Valiente, J. A., Chen, Z. Z., Munoz-Merino, P. J., & Pritchard, D. E. (2017). Copying@Scale: Using harvesting accounts for collecting correct answers in a MOOCs. *Computers & Education*, 108, 96-114. doi:10.1016/j.compedu.2017.01.015

- Alo The MOOCs moment and the end of reform nso-Mencia, M. E., Alario-Hoyos, C., Maldonado-Mahauad, J., Estevez-Ayres, I., Perez-Sanagustin, M., & Kloos, C. D. (2019). Self-regulated learning in MOOCs: Lessons learned from a literature review. *Educational Review*, 71(2), 1-27. doi:10.1080/00131911.2019.1566208
- Ashenafi, M. M., Riccardi, G., & Ronchetti, M. (2015). Predicting students' final exam scores from their course activities. *Proceedings of the 2015 IEEE Frontiers in Education Conference (FIE)*. Texas, USA. doi:10.1109/FIE.2015.7344081
- Azevedo, R., & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of Educational Psychology*, 96(3), 523-535. doi:10.1037/0022-0663.96.3.523
- Bady, A. (2013). The MOOCs moment and the end of reform. *Liberal Education*, 99(4), 6-15.
- Bol, L., & Garner, J. K. (2011). Challenges in supporting self-regulation in distance education environments. *Journal of Computing in Higher Education*, 23(2), 104-123. doi:10.1007/s12528-011-9046-7
- Boulay, B. d. (2016). Artificial intelligence as an effective classroom assistant. *IEEE Intelligent Systems*, 31(6), 76-81. doi:10.1109/MIS.2016.93
- Chen, X., Xie, H., & Hwang, G. J. (2020). A Multi-perspective study on artificial intelligence in education: Grants, conferences, journals, software tools, institutions, and researchers. *Computers and Education: Artificial Intelligence*, 1, 100005. doi:10.1016/j.caeai.2020.100005
- Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 1, 100002. doi:10.1016/j.caeai.2020.100002
- Chu, L., Li, P. H., & Yu, M. N. (2020). The Longitudinal effect of children's self-regulated learning on reading habits and well-being. *International Journal of Educational Research*, 104, 101673. doi:10.1016/j.ijer.2020.101673
- Crosslin, M. (2018). Exploring self-regulated learning choices in a customizable learning pathway MOOCs. *Australasian Journal of Educational Technology*, 34(1), 131-144. doi:131-144. 10.14742/ajet.3758
- Daghestani, L. F., Ibrahim, L. F., Al-Towirgi, R. S., & Salman, H. A. (2020). Adapting gamified learning systems using educational data mining techniques. *Computer Applications in Engineering Education*, 28(3), 568-589. doi:10.1002/cae.22227
- DeBoer, J., Ho, A. D., Stump, G. S., & Breslow, L. (2014). Changing "course" reconceptualizing educational variables for massive open online courses. *Educational Researcher*, 43(2), 74-84. doi:10.3102/0013189X14523038
- Er, E., Gomez-Sanchez, E., Dimitriadis, Y., Bote-Lorenzo, M. L., Asensio-Perez, J. I., & Alvarez-Alvarez, S. (2019). Aligning learning design and learning analytics through instructor involvement: A MOOCs case study. *Interactive Learning Environments*, 27(5-6), 685-698. doi:10.1080/10494820.2019.1610455
- Fauvel, S., Yu, H., Miao, C., Cui, L., Song, H., Zhang, L., Li, X., & Leung, C. (2018). Artificial intelligence powered MOOCs: A Brief survey. *Proceedings of the 2018 IEEE International Conference on Agents (ICA'18)* (pp. 56-61). doi:10.1109/AGENTS.2018.8460059
- Freitas, S. I., Morgan, J., & Gibson, D. (2015). Will MOOCs transform learning and teaching in higher education? Engagement and course retention in online learning provision. *British Journal of Educational Technology*, 46(3), 455-471. doi: 10.1111/bjet.12268
- Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553), 452-459.
- Guo, P. J., Kim, J., & Rubin, R. (2014). How video production affects student engagement: An Empirical study of MOOCs videos. *Proceedings of the first ACM conference on Learning @ scale conference (L@S '14)* (pp. 41-50). New York, NY: ACM. doi:10.1145/2556325.2566239
- Hsu, T. C., Abelson, H., Lao, N., Tseng, Y. H., & Lin, Y. T. (2021). Behavioral-pattern exploration and development of an instructional tool for young children to learn AI. *Computers and Education: Artificial Intelligence*, 2, 1-14. doi: 10.1016/j.caeai.2021.100012
- Huang, N. F., Chen, C. C., Tzeng, J. W., & Lee, C. A. (2018). Concept assessment system integrated with knowledge map using deep learning. *Proceedings of the IEEE Learning with MOOCs (LWMOOCs 2018)* (pp. 113-116), Madrid, Spain.. doi:10.1109/LWMOOCs.2018.8534674
- Hwang, G. J., Chu, H. C., & Yin, C. (2017). Objectives, methodologies and research issues of learning analytics. *Interactive Learning Environments*, 25(2), 143-146. doi: 10.1080/10494820.2017.1287338
- Hwang, G. J., Sung, H. Y., Chang, S. C., & Huang, X. C. (2020). A Fuzzy expert system-based adaptive learning approach to improving students' learning performances by considering affective and cognitive factors. *Computers and Education: Artificial Intelligence*, 1, 1-15. doi:10.1016/j.caeai.2020.100003
- Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 1, 100001. doi:10.1016/j.caeai.2020.100001

- Kastrati, Z., Imran, A. S., & Kurti, A. (2019). Integrating word embeddings and document topics with deep learning in a video classification framework. *Pattern Recognition Letters*, 128, 85-92. doi:10.1016/j.patrec.2019.08.019
- Kavitha, G., & Raj, L. (2017). Educational data mining and learning analytics - Educational assistance for teaching and learning. *International Journal of Computer & Organization Trends*, 41(1), 21-25. doi:10.14445/22492593/IJCOT-V41P304
- Kay, J., & Kummerfeld, B. (2019). From data to personal user models for life-long, life-wide learners. *British Journal of Educational Technology*, 50(6), 2871-2884. doi: 10.1111/bjet.12878
- Kim, J., Guo, P. J., Seaton, D. T., Mitros, P., Gajos, K. Z., & Miller, R. C. (2014). Understanding in-video dropouts and interaction peaks in online lecture videos. *Proceedings of the first ACM conference on Learning @ scale conference (L@S '14)* (pp. 31-40). New York, NY: ACM. doi:10.1145/2556325.2566237
- Kim, R., Olfman, L., Ryan, T., & Eryilmaz, E. (2014). Leveraging a personalized system to improve self-directed learning in online educational environments. *Computers & Education*, 70, 150-160. doi:10.1016/j.compedu.2013.08.006
- Lan, M., Hou, X. Y., Qi, X. Y., & Mattheos, N. (2019). Self-regulated learning strategies in world's first MOOCs in implant dentistry. *European Journal of Dental Education*, 23(3), 278-285. doi:10.1111/eje.12428
- Lee, Y. (2018). Effect of uninterrupted Time-On-Task on students' success in massive open online courses (MOOCs). *Computers in Human Behavior*, 86, 174-180. doi:10.1016/j.chb.2018.04.043
- Lee, Y. (2019). Using self-organizing map and clustering to investigate problem-solving patterns in the Massive Open Online Courses: An Exploratory study. *Journal of Educational Computing Research*, 57(2), 471-490. doi:10.1177/0735633117753364
- Li, C., & Zhou, H. (2018). Enhancing the efficiency of massive online learning by integrating intelligent analysis into MOOCs with an application to education of sustainability. *Sustainability*, 10(2), 468-484. doi:10.3390/su10020468
- Li, K. (2019). MOOCs learners' demographics, self-regulated learning strategy, perceived learning and satisfaction: A Structural equation modeling approach. *Computers & Education*, 132, 16-30. doi:10.1016/j.compedu.2019.01.003
- Li, K.C., & Wong, B. (2020). The Use of student response systems with learning analytics: A Review of case studies (2008-2017). *International Journal of Mobile Learning and Organisation*, 14(1), 63-79. doi:10.1504/IJMLO.2020.103901
- Liu, T. C., Lin, Y. C., & Tsai, C. C. (2009). Identifying misconceptions about statistical correlation and their possible causes among high school students: An Exploratory study using concept mapping with interviews. *International Journal of Science and Mathematics Education*, 7(4), 791-820. doi:10.1007/s10763-008-9142-y
- Li, X., Xie, L., & Wang, H. (2016). Grade prediction in MOOC. *Proceedings of the 2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES)*, Paris, France. doi:10.1109/CSE-EUC-DCABES.2016.213
- Lu, O., Huang, A., Huang, J., Lin, A., Ogata, H., & Yang, S. (2018). Applying learning analytics for the early prediction of students' academic performance in blended learning. *Educational Technology & Society*, 21(2), 220-232.
- Lu, O., Huang, J., Huang, A., & Yang, S. (2017). Applying learning analytics for improving students engagement and learning outcomes in a MOOCs enabled collaborative programming course. *Interactive Learning Environments*, 25(2), 220-234. doi:10.4324/9780429428500-7
- Matt, C. (2018). Exploring self-regulated learning choices in a customisable learning pathway MOOC. *Australasian Journal of Educational Technology*, 34(1), 131-144. doi:10.14742/ajet.3758
- Meier, Y., Xu, J., Atan, O., & Schaar, M. v. d. (2016). Predicting grades. *IEEE Transactions on Signal Processing*, 64(4), 959-972. doi:10.1109/TSP.2015.2496278
- Moreno-Marcos, P. M., Pong, T., Muñoz-Merino, P. J., & Delgado Kloos, C. (2020). Analysis of the factors influencing learners' performance prediction with learning analytics. *IEEE Access*, 8, 5264-5282. doi:10.1109/ACCESS.2019.2963503
- Muñoz-Merino, P. J., Ruipérez-Valiente, J. A., Alario-Hoyos, C., Pérez-Sanagustín, M., & Kloos, C. D. (2015). Precise Effectiveness Strategy for analyzing the effectiveness of students with educational resources and activities in MOOCs. *Computers in Human Behavior*, 47, 108-118. doi:10.1016/j.chb.2014.10.003
- Ndukwe, I.G., & Daniel, B.K. (2020). Teaching analytics, value and tools for teacher data literacy: A Systematic and tripartite approach. *International Journal of Educational Technology in Higher Education*, 17(22), 1-31. doi:10.1186/s41239-020-00201-6
- Peverly, S., Brobst, K., Graham, M., & Shaw, R. (2003). College adults are not good at self-regulation: A Study on the relationship of self-regulation, note taking, and test taking. *Journal of Educational Psychology*, 95, 335-346. doi:10.1037/0022-0663.95.2.335
- Romero, C., & Ventura, S. (2017). Educational data science in massive open online courses. *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery*, 7(1), 1-12. doi:10.1002/widm.1187

- Ruipérez-Valiente, José A., Merino, Pedro. J., Pijera Díaz, H. J., Santofimia, R., & Delgado-Kloos, Carlos. (2017). Evaluation of a learning analytics application for open edX platform. *Computer Science and Information Systems*, 14, 43-43. doi:10.2298/CSIS160331043R
- Schwendimann, B. A. (2017). Perceiving learning at a glance: A Systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies*, 10(1), 30-41. doi:10.1109/TLT.2016.2599522
- Shepard, L. (2001). *The Role of classroom assessment in teaching and learning*. CSE Technical Report. Los Angeles, USA: National Center for Research on Evaluation, Standards, and Student Testing.
- Su, Y. S. & Wu, S. Y. (2021). Applying data mining techniques to explore users behaviors and viewing video patterns in converged IT environments. *Journal of Ambient Intelligence and Humanized Computing*. doi:10.1007/s12652-020-02712-6
- Su, Y. S., Ding, T. J., & Chen, M. Y. (2021). Deep learning methods in internet of medical things for valvular heart disease screening system. *IEEE Internet of Things Journal*. doi:10.1109/JIOT.2021.3053420
- Su, Y. S., Ni, C. F., Li, W. C., Lee, I. H., & Lin, C. P. (2020). Applying deep learning algorithms to enhance simulations of large-scale groundwater flow in IoTs. *Applied Soft Computing*, 92, 106298. doi:10.1016/j.asoc.2020.106298
- Su, Y. S., Suen, H. Y., & Hung, K. E. (2021). Predicting behavioral competencies automatically from facial expressions in real-time video recorded interviews. *Journal of Real-Time Image Processing*. doi:10.1007/s11554-021-01071-5
- Su, Y. S., Chou, C. H., Chu, Y. L., & Yang, Z. F. (2019). A Finger-worn device for exploring chinese printed text with using CNN algorithm on a micro IoT processor. *IEEE ACCESS*, 7, 116529-116541. doi:10.1109/ACCESS.2019.2936143
- Su, Y. S. & Lai, C. F. (2021). Applying educational data mining to explore viewing behaviors and performance with flipped classrooms on the social media platform Facebook. *Frontiers in Psychology*, 12. doi:10.3389/fpsyg.2021.653018
- Sun, D., Mao, Y., Du, J., Xu, P., Zheng, Q., & Sun, H. (2019). Deep learning for dropout prediction in MOOCs. *Proceedings of the 2019 Eighth International Conference on Educational Innovation through Technology (EITT)* (pp. 87-90). Mississippi, USA.
- Sun, Y. Q., Ni, L. H., Zhao, Y. M., Shen, X. L., & Wang, N. (2019). Understanding students' engagement in MOOC: An integration of self-determination theory and theory of relationship quality. *British Journal of Educational Technology*, 50(6), 3156-3174. doi:10.1111/bjet.12724
- Tekin, C., Braun, J., & Schaar, M. v. d. (2015). eTutor: Online learning for personalized education. *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5545-5549), Brisbane, Australia. doi:10.1109/ICASSP.2015.7179032
- Vieira, C., Parsons, P., & Byrd, V. (2018). Visual learning analytics of educational data: A Systematic literature review and research agenda. *Computers & Education*, 122, 119-135. doi:10.1016/j.compedu.2018.03.018
- 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), 716-728. doi:10.1109/JSTSP.2017.2700227

## Automatic Generation of Cloze Items for Repeated Testing to Improve Reading Comprehension

Albert C. M. Yang<sup>1\*</sup>, Irene Y. L. Chen<sup>2</sup>, Brendan Flanagan<sup>3</sup> and Hiroaki Ogata<sup>3</sup>

<sup>1</sup>Graduate School of Informatics, Kyoto University, Japan // <sup>2</sup>Department of Accounting, National Changhua University of Education, Taiwan // <sup>3</sup>Academic Center for Computing and Media Studies, Kyoto University, Japan // yang.ming.35e@st.kyoto-u.ac.jp // irene@cc.ncue.edu.tw // flanagan.brendanjohn.4n@kyoto-u.ac.jp // hiroaki.ogata@gmail.com

\*Corresponding author

**ABSTRACT:** Reviewing learned knowledge is critical in the learning process. Testing the learning content instead of restudying, which is known as the testing effect, has been demonstrated to be an effective review strategy. However, education research recommends that instructors generate practice tests, but this burdens teachers and may also hinder teaching quality. To resolve this issue, the current study applied a modern artificial intelligence technique (BERT) to automate the generation of tests and evaluate the testing effect through e-books in a university lecture ( $N = 74$ ). The last 5 minutes of each course session were utilized to review the taught content by having students either answer cloze item questions or restudy the summary of the core concepts covered in the lecture. A reading comprehension pretest was conducted before the experiment to ensure that the differences in prior knowledge were nonsignificant between groups, and a posttest was performed to examine the effectiveness of testing. In addition, we evaluated students' reading skills and reading engagement through their ability to identify key concepts and their interaction with e-books, respectively. A positive effect was observed for students who engaged in cloze item practice before the end of each class. The results indicated that the repeated testing group exhibited significantly better reading skills and engaged more with e-books than the restudying group did. More importantly, compared with only restudying the key concepts, answering the cloze items questions significantly improved students' reading comprehension. Our results suggest that machine-generated cloze testing may benefit learning in higher education.

**Keywords:** Modern AI, Repeated testing, Testing effect, Test-enhanced learning

### 1. Introduction

Artificial intelligence (AI) refers to “computers that mimic cognitive functions that humans associate with the human mind, such as learning and problem-solving” (Russell & Norvig, 2005, p. 2). With the increasing development of information technologies, AI has been extensively applied in the area of education. For example, Junco and Clem (2015) applied a hierarchical linear regression model to predict the course GPA of students on the basis of their reading engagement. Süzen et al. (2020) combined data mining techniques and clustering to automatically grade short-answer assignments and provide feedback to students. Recently, modern AI has been generally referred to as deep neural network (DNN)-based approaches (Yosinski et al., 2014), and these have been applied in academic fields. Zhang et al. (2019) applied a long short-term memory neural network to build a model that could learn word sequence information, thus enabling it to automatically grade semi-open-ended questions. Furthermore, Okubo et al. (2017) proposed a recurrent neural network (RNN) model to predict the course grade of students using the log data collected by learning management systems. Their results indicated that RNN outperformed other regression models in the prediction tasks.

Repeated testing has been demonstrated to be effective for improving both short-term and long-term memory (Wiklund-Hörnqvist et al., 2014). Although the majority of the positive effects of repeated testing have been identified in laboratory settings (Karpicke, 2017; Rowland, 2014), researchers and practitioners have recently started implementing testing in educational contexts. Greving and Richter (2018) had college students review lecture content 10 minutes before the end of each class and determined that students who reviewed the content by answering short-answer questions performed better than those who answered multiple-choice questions or restudied the summaries of the lecture content in a later retention test. However, the testing questions in a majority of previous studies were created by humans, and creating a practice test for all learning materials is resource intensive. This is typically the case in colleges because many instructors choose to organize their materials on their own instead of using existing textbooks. To address this issue, Mouri et al. (2019) utilized the digital textbook logs of students to automatically generate a personalized quiz for the purpose of reviewing. Olney et al. (2017) applied natural language processing (NLP) techniques to generate cloze item practice tests, and they found the effectiveness of machine-generated and human-generated tests to be comparable. In the domain of modern AI, researchers have begun applying modern AI-based techniques to automatically generate



questions using sentences from texts. Du et al. (2017) introduced an attention-based, sequence-to-sequence model for this task, and the results suggested that their model significantly outperformed the state-of-the-art rule-based system. Moreover, Chan and Fan proposed (2019) a recurrent BERT-based model to perform the task of short-answer question generation. Their model resolved the shortcomings of directly using BERT for text generation. However, the majority of previous studies that applied modern AI techniques were focusing on short-answer question generation. Drawing from those studies, we developed a BERT-based system to automatically generate cloze items for practice and examined whether cloze item practice generated by modern AI techniques produces testing effect and whether it has a positive impact on reading comprehension. Furthermore, we collected students' reading logs to evaluate their reading skills and reading engagement. Our hypothesis was that students' reading skills and reading engagement are improved through repeated testing.

## **2. Literature review**

### **2.1. Reading skills**

Reading skills refer to the ability to understand and recall reading content (Memory, 1983). High-skill readers tend to apply different strategies to extract relevant information from the target content and to better recall learned knowledge during the review stage. This phenomenon occurs more frequently in college because college textbooks often contain longer and more difficult sentences; for many students, such reading demands considerable attention to fully understand the content. Therefore, students with high-level reading skill are expected to perform better than those with weaker skills. Furthermore, a previous study demonstrated that high-skill readers are more likely to comprehend learning content than low-skill readers are (Lorch & Pugzles-Lorch, 1985); thus, low-skill readers seem to face difficulties in identifying the most relevant information in the texts that they read. In support of this claim, other researchers have observed that high-skill and low-skill readers differ in terms of the concepts they perceive to be important in a text (Winograd, 1984). Furthermore, Coiro (2011) indicated that differences in prior knowledge can even be compensated for by adolescents with high reading skills when they are learning with others with prior knowledge.

Text marking is a common and effective reading skill. By highlighting or underlining the most relevant information in a text, students can separate the identified valuable information from other irrelevant content and can easily recall key information during later review. Research has indicated that students who used the highlighting feature in digital textbooks achieved better academic outcomes (Junco & Clem, 2015). However, without considering the content of marked text, students might overuse this skill by simply marking more text. Bell and Limber (2009) indicated that text-marking skills represent students' ability to identify and isolate key information and found that low-skill readers tend to highlight more than high-skill readers do because of their inability to identify relevant concepts. That is, the highlight frequency and reading skills are positively correlated only up to a certain extent—when students are unable to distinguish between critical and trivial textbook content, they may overuse the highlighting strategy. Therefore, the measurement of students' text-marking ability in this study was measured by the content of text they marked, instead of the number of highlights they added. Furthermore, Yue et al. (2015) proposed that the effectiveness of highlighting can be optimized when students are trained on how to use this skill. Therefore, we want to investigate whether students' reading skills can be enhanced by taking practice tests since the questions in tests are the key concepts in materials. We measured students' text-marking ability in e-books to evaluate their reading skills in this study.

### **2.2. Reading engagement**

Reading has been shown to directly correlate with course outcomes (Daniel & Woody, 2013). Landrum, Gurung, and Spann (2012) observed that students' self-reported percentage of completed readings in textbooks strongly related to their quiz scores and final grades. Junco and Clem (2015) collected students' engagement index to predict their course outcomes. They found that the time spent on reading was the most significant factor in their prediction model. In addition, reading engagement was found to vary for different texts, with more advanced lectures requiring more reading time (Fitzpatrick & McConnell, 2009). Studies have highlighted that although many students may not read a complete text, they do engage with the interactive features in digital textbooks, and such engagement improves their learning outcomes (Berry et al., 2010; Dennis, 2011; Fouh et al., 2014). Dennis (2011) discovered that the number of annotations was positively related to learning outcomes, whereas the number of pages students read was not, which seems to contradict the finding of Junco and Clem (2015). This suggests that it might not be enough to measure students' reading engagement solely by reading time or the number of pages read; instead, annotation tools, including notes or highlights, allow student to interact with the

text and, in turn, reflect the effort they make during reading, should be considered as well. Therefore, textbook analytics can be applied to measure reading engagement with e-books, and this indicator can be employed to predict students' learning outcomes (Bossaller & Kammer, 2014). In support of this claim, research has demonstrated that students who read more or interact more with their textbooks perform better in class (Dawson, McWilliam & Tan, 2008; DeBerard, Speilmans, & Julka, 2004; Woody et al., 2010). In sum, improving students' reading engagement not only motivates them to interact with the text but also improves their learning performance. Testing has been shown to improve students' learning engagement as they need to spend more time on reading textbooks and readjust their learning strategies in order to answer the questions (Soderstrom & Bjork, 2014). In this study, we measured students' reading engagement by both reading time and the number of annotation tools they used and hypothesized that students' reading engagement with digital textbooks increases after appearing for cloze test practice.

### 2.3. Repeated testing

Traditionally, testing is used to assess students' knowledge and assign grades. However, its employment to facilitate learning is an application of testing that has been largely neglected by educationalists (Butler & Roediger, 2007). Empirical studies have emphasized that compared with traditional restudying of learning materials, taking repeated tests greatly improves students' performance in later recall tests (Butler & Roediger, 2007; McDaniel et al., 2007). One explanation for this effect is that repeated testing forces student to reencode the information they have learned, whereas restudying requires them to only reproduce the encoding of the learned knowledge (Karpicke & Roediger, 2008). The superiority of repeated testing over restudying learning material is known as the testing effect (Roediger & Karpicke, 2006a). Compared with simply rereading the learning material, students subjected to quizzes after reading a chapter of a textbook or upon completion of a course exhibited improved long-term retention of knowledge. This phenomenon is known as the direct testing effect. The indirect testing effect refers to the use of improved strategies or increased motivation to study in anticipation of taking a test. Soderstrom and Bjork (2014)'s results revealed that practice testing motivated participants to readjust their monitoring process and therefore enhanced their learning engagement. Recently, studies on testing effects have gradually shifted from laboratory settings to real classrooms. Bobby et al. (2018) reported that the testing effect of a closed book examination combined with feedback was effective in improving the learning performances for medical students studying biochemistry. Schwierien et al. (2017) conducted a meta-analysis of testing effect and identified a significant overall effect size of  $d = 0.56$ , highlighting that testing was beneficial to the learning outcomes of psychology students. The number of tests a student can take during the practice phase is a key aspect of the testing effect. Repeated testing has been demonstrated to improve retention as opposed to a singular test (Karpicke & Roediger, 2008). Moreover, the effects of repeated testing are more pronounced when tests are administered over time (Karpicke & Roediger, 2007). Another crucial aspect is the provision of feedback. Feedback enhances the benefit of testing through the correction of errors and confirmation of correct answers (Butler & Roediger, 2008). Studies have demonstrated that feedback can dramatically amplify the knowledge retention achieved through repeated testing (Butler et al., 2008). Generally, testing effects are larger for more difficult tests because they require more cognitive effort for information retrieval (Kang et al., 2007). However, raising the difficulty level of tests may lead to increased unsuccessful retrieval. According to one study, retrieval must be successful to reap the benefits of repeated testing (Rowland, 2014). Therefore, feedback can be useful for overcoming the limited effect of unsuccessful retrieval by correcting incorrect responses (Rowland, 2014). Wiklund-Hörnqvist et al. (2014) demonstrated that compared with short- and long-term restudying, repeated testing with feedback significantly promoted learning. Furthermore, they emphasized the importance of educationalists adopting teaching methods that involve repeated testing. With the advancement of information technology, researchers have started applying AI in repeated testing by automatically generating practice tests. For example, Olney et al. (2017) applied NLP techniques to automatically generate cloze items and found machine-generated items to be as effective as human-generated ones for enhancing reading comprehension. In this study, we hypothesized that the direct testing effect will promote student retention of learned knowledge and therefore achieve better scores in the reading comprehension posttest. In addition, we hypothesized that students' reading engagement and reading skills will be enhanced by readjusting their reading behaviors after practice testing. We leveraged modern AI techniques to automatically generate cloze item practice for repeated testing and addressed the following research questions:

- (1) Can students improve their reading skills with machine-generated cloze item practice?
- (2) Can students improve their reading engagement with machine-generated cloze item practice?
- (3) Can students improve their reading comprehension with machine-generated cloze item practice?

### 3. Methods

#### 3.1. Research context

A 4-week experiment was conducted in two mandatory courses for undergraduate students from the accounting department at a university in Taiwan. These courses could be taken as elective courses by students from other departments as well. Both classes were taught by the same instructor using the same materials. A total of 74 students enrolled in this experiment. Both courses employed BookRoll, an e-book reading system (Flanagan & Ogata, 2017) developed by Kyoto University; instructors can upload materials, and students can use the e-book reader to read the content and interact with the text using the provided tools, such as notes and highlights. The actions performed by students are stored in the database for later analysis. The e-book reading actions available in BookRoll have been described in detail by Ogata et al. (2015) and Flanagan and Ogata (2018). Participants took a reading comprehension pretest and posttest during the first and the final week of the experiment that evaluated whether the use of cloze item practice promoted their reading proficiency. The reading comprehension pretest and posttest each comprised 28 multiple-choice questions that had been randomly extracted from a test bank with 50 questions related to the accounting field. The test bank had been created by two instructors at the department with accounting experience.

#### 3.2. Procedure

During the experiment, one class was assigned to be the experimental group and the other constituted the control group. In the first week of the experiment, students were asked to complete a reading comprehension pretest. The instructor uploaded the materials a week before each class. Students were required to review the materials and mark the sentences or words that they thought were important. Their marking scores were calculated according to the content they had marked, which was considered to be a reflection of their reading skills. Moreover, the actions students performed during their reading were examined to assess their reading engagement. The measurement of reading skills and reading engagement is explained in the following section. The instructor briefly discussed the content of the materials shared and answered students' questions during the class. Students in the experimental group were required to take a cloze test practice at the end of each class, whereas the control group students restudied the key concepts in the learning materials summarized by our system. To investigate whether different review methods affect learning, the questions in cloze item practice for experimental group and the key concepts for control group consisted of the same sentences extracted from the materials, except that one or two words in each sentence were masked for the questions, whereas the original sentences were presented in key concepts. The experimental group students could take the test and practice (the number of correct answers was not counted in their final course grade) repeatedly. The experimental group students were encouraged to test themselves after class, and the control group were encouraged to restudy the key concepts as well. Finally, both groups took a reading comprehension posttest in the last week of the experiment, and the results were used to evaluate the effectiveness of cloze item practice. The questions in the reading comprehension test were different from the questions in cloze item practice and key concepts presented to students.

#### 3.3. Automatic cloze item generation

We applied the advanced neural network technique BERT and the machine learning model TextRank to generate cloze items in this study. BERT is a pretrained model that was developed by Google for NLP. During the pretraining phase, BERT develops bidirectional representations from a plain text corpus by taking into account the context of each occurrence of a given word. Thus, unlike other word-embedding models such as Word2vec or GloVe that create a single-word embedding for each word, BERT generates a contextualized embedding representation that varies depending on the sentence. As a result, the pretrained model can be fine-tuned by simply introducing an additional layer to create a specific model for various tasks such as question answering and language inference. TextRank is an unsupervised machine learning algorithm based on PageRank, which is often used for keyword extraction and text summarization. TextRank constructs a graph denoting the relationships between the words in a text and ranks the items in the graph. This method allows TextRank to generate summaries without a training corpus or labeling and makes it appropriate for application in various language tasks. In this study, the open-source transformers packages developed by Hugging Face and TextRank4ZH were adopted to implement BERT and TextRank, respectively.

In our study, the generation of cloze items involved two steps: key sentence extraction and keyword extraction. First, we split the text into sentences and applied BERT to generate the embedding of the full text and the

embedding of each sentence. The cosine distance between the embedding of the text and the embedding of each sentence was calculated, and the sentences that were close to the text in the vector space were selected as the core concepts. The selected sentences were provided to the control group students to review. Second, TextRank was applied to extract keywords from each selected sentence. Subsequently, words with the highest weight were masked as cloze items for the experimental group. Figure 1 shows a snapshot of cloze item practice. When students enter the module, they need to choose the e-Book they want to review. The class name, e-Book, and student ID will be displayed. Students are aware of the total number of questions and the number of questions they have completed. When students click a mask, an input field will be displayed. Then, students need to enter and submit their answer. They are not required to answer the questions in order. For example, they can jump between the pages to answer the questions they are familiar with first, or skip the questions that they already know the answer. After completing the practice testing, they close the module to leave the system. All students' behaviors during testing will be recorded in the database for future analysis.

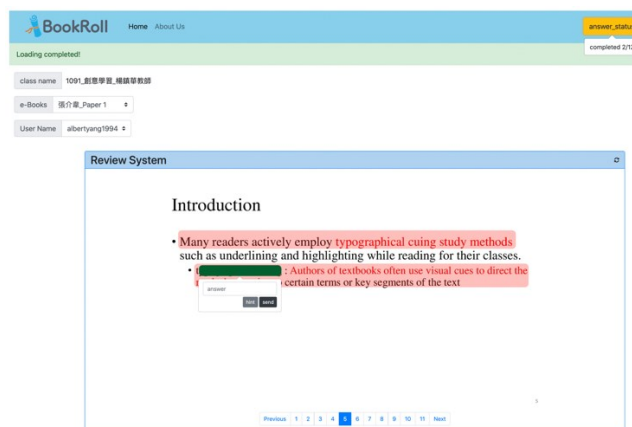


Figure 1. A snapshot of activities in testing module

### 3.4. Measurement of reading skills and reading engagement

According to Bell and Limber (2009), text-marking skills indicate a reader's ability to identify relevant information in a text, and only high-skill readers are able to achieve this task. Hence, we utilized the sentences generated by BERT during the creation of cloze items as essential information in the text, after which we calculated the similarity between those sentences and the content marked by students using Bilingual Evaluation Understudy (BLEU; Papineni et al., 2002). The BLEU score was subsequently employed as the marking score to represent reading skills. The score ranged from 0.0 to 1.0 and was calculated every week. A higher score denoted better reading skills. We used students' reading actions on BookRoll to assess their reading engagement. The actions included their reading time (25%), the number of highlights made (25%), the number of memos posted (25%), and the number of bookmarks added (25%). All feature values were standardized, and the score of their reading engagement was calculated by the sum of the weighted feature values. For example, if student A's standardized score of reading time is 70, standardized score at making highlights is 80, standardized score of posting memos is 60, and the standardized score of adding bookmarks is 60, the reading engagement score of student A is  $70 * 0.25 + 80 * 0.25 + 60 * 0.25 + 60 * 0.25 = 67.5$ . Therefore, more actions on BookRoll indicated higher reading engagement.

## 4. Results

### 4.1. Analysis of Reading Skills and Reading Engagement

An independent  $t$  test was performed to evaluate the influence of cloze item practice on reading skills. The results of the Levene test were not significant ( $F = 0.28, p = .59$ ), indicating that variance homogeneity existed between the groups. As presented in Table 1, the experimental group exhibited a significantly higher marking score than the control group did ( $t = 2.70, p < .01$ ). The mean and standard deviations for the experimental group were 66.34 and 11.21, respectively, and those for the control group were 59.52 and 10.51, respectively. These results suggested that students' reading skills improved after the administration of cloze item practice.

Table 1. Independent *t*-test result of the marking scores of two groups

Dimension	Group	<i>N</i>	Mean	<i>SD</i>	<i>t</i>
Marking score	Experimental group	36	66.34	11.21	2.70**
	Control group	38	59.52	10.51	

Note. \*\*  $p < .01$ .

Subsequently, we measured the differences in reading engagement between the groups. The Levene test for determining the homogeneity of variance showed no violations ( $F = 0.00$ ,  $p = .92$ ), indicating that the assumption was tenable and that the independent *t* test could be used to interpret the relationship between the application of cloze item practice and reading engagement. Table 2 shows that the experimental group exhibited a significantly higher reading engagement than the control group did ( $t = 2.34$ ,  $p < .05$ ). The mean and standard deviations of the experimental group and control group were 75.77 and 11.59 and 69.05 and 13.00, respectively. This indicated that students demonstrated more reading engagement with their e-books after the use of cloze item practice. Furthermore, the independent *t* test was performed again to compare the reading time of two groups outside the class. The Levene test results indicated the homogeneity of variance existed in two groups ( $F = 0.02$ ,  $p = .88$ ). The independent *t* test results showed that experimental group had a significantly higher reading time outside the class than the control group had ( $t = 2.28$ ,  $p < .05$ ; Table 2), meaning that students spent more time on reading after class in order to pass the practice testing. The mean and standard deviations for the experimental group were 455.19 and 370.55, respectively, and those for the control group were 250.00 and 401.79, respectively.

Table 2. Independent *t*-test results of the reading engagement and the reading time outside the class of both groups

Dimension	Group	<i>N</i>	Mean	<i>SD</i>	<i>t</i>
Reading engagement	Experimental group	36	75.77	11.59	2.34*
	Control group	38	69.05	13.00	
Reading time outside class (minutes)	Experimental group	36	455.19	370.55	2.28*
	Control group	38	250.00	401.79	

Note. \*  $p < .05$ .

#### 4.2. Analysis of reading comprehension

After obtaining the pretest and posttest results concerning reading comprehension, we analyzed the mean and standard deviation of the data and used the Python package Pingouin to conduct a one-way analysis of covariance (ANCOVA), where the covariate was the pretest score, the independent variable was the use of cloze item practice, and the dependent variable was the posttest score. The mean and standard deviations of the posttest scores of both groups are presented in Table 3. The pretest and posttest each comprised 28 multiple-choice questions. A total of 28 points could be scored on each test. The *t* test outcome of the pretest was  $t = 1.31$ ,  $p = 0.19$ . This indicated that no significant discrepancy existed between the prior knowledge of both groups.

Table 3. Pretest and posttest scores for reading comprehension under different review conditions

Factors	Control group		Experimental group	
	Mean	<i>SD</i>	Mean	<i>SD</i>
Pretest score				
Reading comprehension	23.68	1.49	24.16	1.65
Posttest score				
Reading comprehension	23.86	1.29	25.50	0.79

One-way ANCOVA was performed to verify whether the between-group differences in the reading comprehension results of the pretest and posttest were statistically significant. Regression coefficients revealed no significant interaction between the covariates and independent variables ( $F = 0.78$ ,  $p = .68$ ); hence, the regression coefficients within the groups did not violate the assumption of homogeneity. Likewise, the results of the Levene test were not significant ( $F = 3.25$ ,  $p = .07$ ). This indicated that homogeneity of variance existed between the groups and that one-way ANCOVA could be conducted to explore any significant differences in the reading comprehension posttest scores of the two groups. The mean of the posttest scores between students in the experimental group (Mean = 25.50, *SD* = 0.79, Adjusted mean = 25.45) and control group (Mean = 23.86, *SD* = 1.29, Adjusted mean = 23.90) was significantly different ( $F = 38.83$ ,  $p < .001$ ,  $\eta^2 = 0.34$ ; Table 4). This finding suggested that students who repeatedly tested themselves showed largely improved reading comprehension compared with those who restudied the materials. Moreover, an independent *t* test was employed to measure the

within-subject difference in the posttest scores. Students in the experimental group exhibited significantly improved performance ( $t = 4.35$ ,  $p < .001$ ), whereas students who restudied the materials failed to exhibit a significant improvement in their posttest scores compared with their pretest scores ( $t = -0.86$ ,  $p = 0.39$ ; Figure 2). Figure 3 presents that both low-skill readers and high-skill readers of the experimental group achieved a better performance in their posttest.

Table 4. Posttest scores for reading comprehension under different review conditions

Source of variance	<i>SS</i>	<i>df</i>	<i>F</i>	$\eta^2$
Covariates	4.85	1	4.33*	0.03
Intergroup	43.47	1	38.83***	0.34
Residual	79.48	71		

Note. \* $p < .01$ ; \*\*\* $p < .001$ .

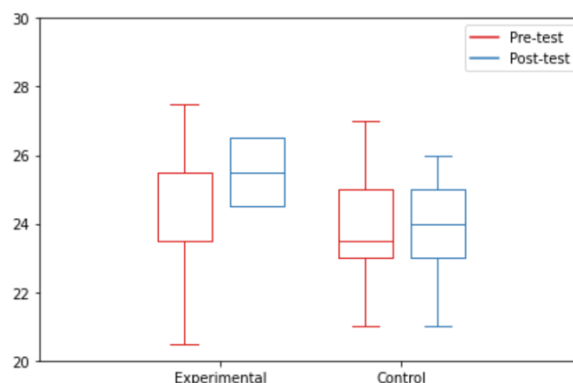


Figure 2. Within-subject differences in the pretest and posttest scores of the experimental group (repeated testing) and the control group (restudying)

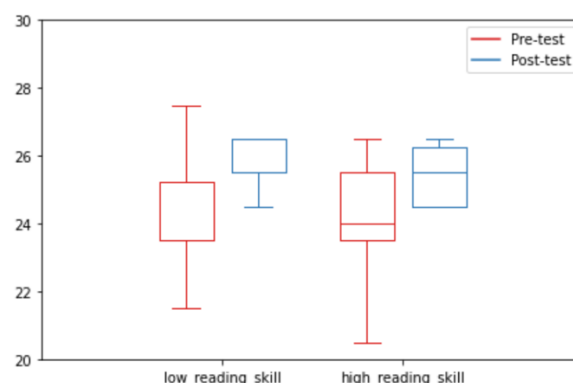


Figure 3. Difference in the pretest and posttest scores of the low-skill readers and high-skill readers in experimental group

## 5. Discussion

### 5.1. Differences in reading skills and reading engagement

#### 5.1.1. Research Question 1: Can students improve their reading skills with machine-generated cloze item practice?

The first question addressed by this study was whether repeated testing had a positive effect on students' reading skills. We assumed that if the students were directly shown and tested on the key concepts of a text, they would better understand the core material of the class and their reading skills could be improved. Our results revealed that students who took the review test achieved a significantly better marking score than those who restudied the materials, indicating that they were superior at finding the key concepts in a text. Nist and Simpson (1988) and Yue et al. (2015) have contended that the effectiveness of marking or underlining can be optimized when students are trained on how to use these skills. Because the test questions generated by our system included key sentences and keywords from the materials, students could compare the sentences in the list of questions with

those that they had highlighted. This process indirectly showed students the correct method to mark the key concepts. The restudy group also reviewed the important concepts in the text; however, repeated testing was found to better promote retention than restudying did (Butler & Roediger, 2007; McDaniel et al., 2007). Therefore, the experimental group demonstrated better reading skills. According to Bell and Limber (2009), high-skill readers are superior at identifying the important information in a text compared with their counterparts. By repeatedly taking the after-class practice tests, students learned how to correctly highlight the important concepts, which, in turn, improved their marking scores and reading skills.

### ***5.1.2. Research question 2: Can students improve their reading engagement through machine-generated cloze item practice?***

We examined whether students who used the reviewing system demonstrated different levels of reading engagement than those who restudied the materials. The results indicated that the experimental group students showed higher reading engagement than the control group did at a statistically significant level, meaning that they spent more time reading the e-books outside the class. To be able to answer the questions in the practice test, the experimental group students needed to review the e-books before taking the test. Therefore, they likely had more reading time than the control group students did and used interactive tools to facilitate their review process. After the test, they adjusted their reading skills on the basis of the results, suggesting that repeated testing motivated them to interact with the materials. Repeated testing can be used as a tool by students to evaluate their reading skills and revise it according to the results. The more tests students take, the more effort they put into learning. These effects are known as the indirect effects of testing (Roediger & Karpicke, 2006a). The increased duration of learning and improved reading skills after taking tests facilitate students' reading engagement and reading comprehension (Olney et al., 2017; Larsen et al., 2009).

## **5.2. Improved reading comprehension**

### ***5.2.1. Research question 3: Can students improve their reading comprehension with machine-generated cloze item practice?***

We explored whether testing promoted students' reading comprehension. The study results revealed that students who took the practice test demonstrated significant improvement in reading comprehension compared with those who restudied the materials. This finding was consistent with the benefit of testing effect highlighted by Wiklund-Hörnqvist et al. (2014), who had also conducted an experiment in which feedback was provided. Meanwhile, the questions in reading comprehension posttest were different from the questions in cloze item practice for experimental group and the key concepts for control group, meaning that the improvement in reading comprehension were not caused by having more opportunities to practice the questions, as two groups were reviewing the same knowledge. Instead, it is the review methods that contributed to the difference in learning performance. Knowledge of key concepts is critical for students to comprehend a course (Kintsch et al., 1998). Studies have shown that learning the meaning of keywords improves reading comprehension (McDaniel & Pressley, 1989). However, different students normally exhibit various levels of reading skills in educational contexts—high-skill readers read more and learn more key concepts than low-skill readers do (Mol & Bus, 2011). In the present study, we addressed this reality by automatically generating practice tests that included key concepts using NLP techniques; we expected to reduce the gap in knowledge concerning key concepts between readers with different reading skills. The results highlighted that students who took the practice test demonstrated improved reading comprehension, regardless of their reading skills. This indicated that students with low reading skills could understand essential information even if they had failed to identify it before the test.

Provision of feedback is another factor that can improve reading comprehension. Kornell et al. (2011) stated that practice without feedback leads to a bifurcated item distribution in which only those items that are successfully retrieved are highly accessible by memory, whereas items that are not retrieved do not result in the testing effect. When students are provided with feedback, their memory strength becomes high enough to exceed a certain threshold; upon this threshold being crossed, the information becomes recallable. This promotes memory retention and prevents erroneous learning. Rowland (2014) indicated that no testing effect can be observed in the absence of feedback and that the retrievable rate is  $\leq 50\%$  in a laboratory setting. In the current study, the mask was removed from the cloze items for correct responses. Furthermore, students were allowed to see a hint if they could not answer correctly, which made each item recallable during every attempt. Our results indicated that the combination of repeated cloze item practice and the provision of feedback engendered the testing effect of enhancing students' reading comprehension.

## 6. Conclusion

Repeated testing has been shown to be an effective strategy for promoting memory retention and learning motivation. In this study, we employed cloze item practice that was automatically generated by BERT to explore two indirect testing effects, namely improvement in reading skills and reading engagement, and one direct testing effect, namely enhancement of reading comprehension. The results indicated that repeated testing significantly improved students' ability to extract key concepts from a text and motivated them to actively read the e-Books before and after the test, respectively. More importantly, their retention of learning content was also enhanced.

Several contributions are made by this study. First, the present study applied a modern AI technique to automatically generate tests for repeated testing in a real educational context. A majority of related studies that have examined the testing effect required instructors to prepare the practice test (Butler & Roediger, 2007; McDaniel et al., 2007; Wiklund-Hörnqvist et al., 2014). Although Olney et al. (2017) proposed a model that automatically generated cloze items for practice, they were using traditional machine learning and NLP techniques. The present study applied a DNN model (BERT) to build a system for generating practice tests. The results indicated that the use of cloze item practice along with the provision of feedback yielded a testing effect that positively influenced reading skills, reading engagement, and reading comprehension. Second, although the benefit of the testing effect has been broadly discussed in many studies (Greving & Richter, 2018; Wiklund-Hörnqvist et al., 2014), most have focused only on the improvement in the retention of taught content. Our study, by contrast, explored whether the testing effect is beneficial for not only students' reading comprehension but also their cognitive behaviors (i.e., reading skills and reading engagement) and determined that test-enhanced learning promotes students' ability to identify important information and motivates them to read. Finally, whether the question format influences the effectiveness of testing has been well investigated in prior research (Greving & Richter, 2018), with findings demonstrating that both short-answer type and multiple-choice questions yield the testing effect; however, the efficacy of other question formats has rarely been discussed. One of our objectives in this study was to investigate whether testing with cloze items is also effective for improving learning. True to our hypothesis, students who took the cloze item practice after class demonstrated greatly improved comprehension.

The current study's findings offer insights for instructors and researchers in related fields. Instructors can use these findings as a reference for guiding students in distinguishing relevant information from trivial content by testing the key concepts and enhancing their reading engagement. Furthermore, the summary generated by our model can be applied in other educational contexts. For example, instructors can use the summary to perform a test before a class to understand the average knowledge level of the class. The instructor can also adjust the summary by adding more sentences that they expect their students to learn or by removing some irrelevant sentences from the summary to develop a personalized summary that closely fits the course objectives. Furthermore, the current study suggests that the automatically generated cloze items are effective in enhancing students' comprehension. Future researchers can apply the same model as our study (BERT) or other modern AI techniques to generate different formats of questions, such as short-answer questions, for repeated learning. Moreover, researchers can develop personalized tests for individual students on the basis of their prior knowledge to improve their learning.

The present study has three limitations that warrant mention. First, the materials used in our experiment concerned topics that involve students' memory (accounting). Although repeated testing has been shown to improve memory retention, whether testing is still effective in promoting learning for materials that require logic and computation is unclear. Second, despite the encouragement given to the experimental group students to use the proposed system outside of class, this action was not mandatory. Therefore, we were unable to evaluate whether repeated testing promotes retention better than taking a single test does (Karpicke & Roediger, 2008). Finally, despite the retention test was conducted at the end of the experiment to measure students' comprehension, which we considered as a relatively long period, it is unclear whether the testing effects promoted long-term or short-term retention as students may make extensive use of the system to review right before appearing for a retention test. In this case, we can only argue that the testing effects in this experiment provided short-term retention.

In sum, our study results demonstrate that testing with BERT-generated cloze items is effective in promoting students' reading skills, reading engagement, and reading comprehension at the undergraduate level. More modern AI-driven testing can be applied to educationally relevant materials to facilitate learning. In our future research, students' review behaviors will be analyzed during testing and a personalized test will be generated on



the basis of their learning profile. Furthermore, we expect to try other DNN models for generating other question formats to develop a more comprehensive test.

## Acknowledgement

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 JPNP20006 and JPNP18013.

## References

- Bell, K. E., & Limber, J. E. (2009). Reading skill, textbook marking, and course performance. *Literacy research and instruction*, 49(1), 56-67.
- Berry, T., Cook, L., Hill, N., & Stevens, K. (2010). An Exploratory analysis of textbook usage and study habits: Misperceptions and barriers to success. *College Teaching*, 59(1), 31-39.
- Bobby, Z., Nandeesha, H., Thippeswamy, D. N., Archana, N., Prerna, S., & Balasubramanian, A. (2018). 'Test-enhanced learning' by Closed book examination followed by feedback in Biochemistry. *South-East Asian Journal of Medical Education*, 12(2), 19-24.
- Butler, A. C., & Roediger III, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, 19(4-5), 514-527.
- Butler, A. C., Karpicke, J. D., & Roediger III, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(4), 918.
- Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & cognition*, 36(3), 604-616.
- Bossaller, J., & Kammer, J. (2014). Faculty views on eTextbooks: A Narrative study. *College Teaching*, 62(2), 68-75.
- Chan, Y. H., & Fan, Y. C. (2019, November). A Recurrent BERT-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering* (pp. 154-162). Hong Kong, China: Association for Computational Linguistics.
- Coiro, J. (2011). Predicting reading comprehension on the Internet: Contributions of offline reading skills, online reading skills, and prior knowledge. *Journal of literacy research*, 43(4), 352-392.
- Daniel, D. B., & Woody, W. D. (2013). E-textbooks at what cost? Performance and use of electronic v. print texts. *Computers & Education*, 62, 18-23.
- Dawson, S. P., McWilliam, E., & Tan, J. P. L. (2008). Teaching smarter: How mining ICT data can inform and improve learning and teaching practice. *Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education* (pp. 221-230). Melbourne, Australia: Deakin University.
- DeBerard, M. S., Spielmans, G. I., & Julka, D. L. (2004). Predictors of academic achievement and retention among college freshmen: A Longitudinal study. *College student journal*, 38(1), 66-81.
- Dennis, A. (2011). e-Textbooks at Indiana University: A Summary of two years of research. *IRB*, 912000863(1003001166), 0908000546. Retrieved from <https://assets.uits.iu.edu/pdf/eText%20Pilot%20Data%202010-2011.pdf>
- Du, X., Shao, J., & Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. Retrieved from <https://arxiv.org/abs/1705.00106>
- Fitzpatrick, L., & McConnell, C. (2009). Student reading strategies and textbook use: An Inquiry into economics and accounting courses. *Research in Higher Education Journal*, 3, 1-10.
- Flanagan, B., & Ogata, H. (2017). Integration of learning analytics research and production systems while protecting privacy. In *Proceedings of the 25th International Conference on Computers in Education* (pp. 333-338). New Zealand: Asia-Pacific Society for Computers in Education.
- Flanagan, B., & Ogata, H. (2018). Learning analytics infrastructure for seamless learning. In *Proceedings of the 8th International Conference on Learning Analytics & Knowledge (LAK18)*. Sydney, Australia: Association for Computing Machinery (ACM).
- Fouh, E., Breakiron, D. A., Hamouda, S., Farghally, M. F., & Shaffer, C. A. (2014). Exploring students learning behavior with an interactive etextbook in computer science courses. *Computers in Human Behavior*, 41, 478-485.
- Greving, S., & Richter, T. (2018). Examining the testing effect in university teaching: Retrieval and question format matter. *Frontiers in Psychology*, 9, 2412.

- Junco, R., & Clem, C. (2015). Predicting course outcomes with digital textbook usage data. *The Internet and Higher Education*, 27, 54-63.
- Karpicke, J. D., & Roediger III, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(4), 704-719. doi:10.1037/0278-7393.33.4.704
- Karpicke, J. D., & Roediger, H. L. (2008). The Critical importance of retrieval for learning. *science*, 319(5865), 966-968.
- Karpicke, J. D. (2017). Retrieval-based learning: a decade of progress. In J. T. Wixted (Ed.), *Cognitive Psychology of Memory, of Learning and Memory: A Comprehensive Reference* (Vol. 2, pp. 487-514). Oxford, UK: Academic Press. doi:10.1016/B978-0-12-809324-5.21055-9
- Kang, S. H., McDermott, K. B., & Roediger III, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, 19(4-5), 528-558.
- Kintsch, W. (1998). *Comprehension: A Paradigm for cognition*. Cambridge, UK: Cambridge University Press.
- Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A Distribution-based bifurcation model. *Journal of Memory and Language*, 65(2), 85-97.
- Landrum, R. E., Gurung, R. A., & Spann, N. (2012). Assessments of textbook usage and the relationship to student course performance. *College Teaching*, 60(1), 17-24.
- Larsen, D. P., Butler, A. C., & Roediger III, H. L. (2009). Repeated testing improves long-term retention relative to repeated study: A Randomised controlled trial. *Medical education*, 43(12), 1174-1181.
- Lorch, R. F., & Lorch, E. P. (1985). Topic structure representation and text recall. *Journal of Educational Psychology*, 77(2), 137-148. doi:10.1037/0022-0663.77.2.137
- McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, 19(4-5), 494-513.
- McDaniel, M. A., & Pressley, M. (1989). Keyword and context instruction of new vocabulary meanings: Effects on text comprehension and memory. *Journal of Educational Psychology*, 81(2), 204-213. doi:10.1037/0022-0663.81.2.204
- Memory, D. M. (1983). Main idea prequestions as adjunct aids with good and low-average middle grade readers. *Journal of Reading Behavior*, 15(2), 37-48.
- Mol, S. E., & Bus, A. G. (2011). To read or not to read: A Meta-analysis of print exposure from infancy to early adulthood. *Psychological bulletin*, 137(2), 267-296. doi:10.1037/a0021890.
- Mouri, K., Uosaki, N., Hasnine, M., Shimada, A., Yin, C., Kaneko, K., & Ogata, H. (2019). An Automatic quiz generation system utilizing digital textbook logs. *Interactive Learning Environments*, 1-14. doi:10.1080/10494820.2019.1620291
- Nist, S. L., & Simpson, M. L. (1988). The effectiveness and efficiency of training college students to annotate and underline text. *National Reading Conference Yearbook*, 37, 251-257.
- Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015). E-Book-based learning analytics in university education. In *International Conference on Computer in Education (ICCE 2015)* (pp. 401-406). China: Asia-Pacific Society for Computers in Education
- Okubo, F., Yamashita, T., Shimada, A., & Ogata, H. (2017). A Neural network approach for students' performance prediction. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 598-599). New York, NY: Association for Computing Machinery.
- Olney, A. M., Pavlik, P. I., & Maass, J. K. (2017). Improving reading comprehension with automatically generated cloze item practice. In *Artificial intelligence in education* (pp. 262-273). Cham: Springer International Publishing. doi:10.1007/978-3-319-61425-0\_22
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A Method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318). Philadelphia, PA: Association for Computational Linguistics.
- Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, 17(3), 249-255.
- Rowland, C. A. (2014). The Effect of testing versus restudy on retention: Meta-analytic review of the testing effect. *Psychological Bulletin*, 140(6), 1432-1463. doi:10.1037/a0037559
- Russell, S., & Norvig, P. (2005). AI a modern approach. *Learning*, 2(3), 4.
- Schwieren, J., Barenberg, J., & Dutke, S. (2017). The Testing effect in the psychology classroom: A Meta-analytic perspective. *Psychology Learning & Teaching*, 16(2), 179-196.

- Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language*, 73, 99-115.
- Süzen, N., Gorban, A. N., Levesley, J., & Mirkes, E. M. (2020). Automatic short answer grading and feedback using text mining methods. *Procedia Computer Science*, 169, 726-743.
- Wiklund-Hörnqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian journal of psychology*, 55(1), 10-16.
- Winograd, P. N. (1984). Strategic difficulties in summarizing texts. *Reading Research Quarterly*, 19(4), 404-425.
- Woody, W. D., Daniel, D. B., & Baker, C. A. (2010). E-books or textbooks: Students prefer textbooks. *Computers & Education*, 55(3), 945-948.
- Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems* (Vol. 27, pp. 3320-3328). Retrieved from <https://arxiv.org/abs/1411.1792>
- Zhang, L., Huang, Y., Yang, X., Yu, S., & Zhuang, F. (2019). An automatic short-answer grading model for semi-open-ended questions. *Interactive Learning Environments*, 1-14. doi:10.1080/10494820.2019.1648300

## Expert-Authored and Machine-Generated Short-Answer Questions for Assessing Students' Learning Performance

Owen H. T. Lu<sup>1</sup>, Anna Y. Q. Huang<sup>2</sup>, Danny C. L. Tsai<sup>2</sup> and Stephen J. H. Yang<sup>2\*</sup>

<sup>1</sup>College of Computer Science, National Pingtung University, Taiwan // <sup>2</sup>Department of Computer Science and Information Engineering, National Central University, Taiwan // owen.lu.academic@gmail.com // anna.yuqing@gmail.com // dan860202@gmail.com // stephen.yang.ac@gmail.com

\*Corresponding author

**ABSTRACT:** Human-guided machine learning can improve computing intelligence, and it can accurately assist humans in various tasks. In education research, artificial intelligence (AI) is applicable in many situations, such as predicting students' learning paths and strategies. In this study, we explore the benefits of repetitive practice of short-answer questions could enhance students' long-term memory for subsequent improvements in learning performance. However, frequent authoring questions and grading requires teachers' professionalism, experience, and considerable efforts. Therefore, this study using modern AI technologies, specifically natural language processing, to provide Automatic question generation (AQG) solution, a combined semantics-based and syntax-based question generation system: Hybrid automatic question generation (Hybrid-AQG) was proposed in this study. We assessed its functionality and student learning performance by asking 91 students to complete short-answer questions and then applied a process similar to the Turing test to evaluate the question and grading quality. The results demonstrated that modern AI technologies can generate highly realistic short-answer questions because: (1) Compared with the control group, the experimental group exhibited significantly better learning performance, implying that students acquired long-term memory of course knowledge through repetitive practice with machine-generated questioning. (2) The experimental group could better distinguish machine-generated and expert-authored questions. Nevertheless, both groups in distinguishing questions presented like guessing. (3) Machine grading was deficient in some respects; but the way students answer questions can be adapted for machine understanding through repetitive practice.

**Keywords:** Automatic question generation, Learning performance, Artificial intelligence, Practice testing, Turing test

### 1. Introduction

#### 1.1. Artificial Intelligence in Education (AIED)

Artificial intelligence (AI) refers to machines thinking and acting rationally like humans (Russell & Norvig, 2002). One of the AI implementation is the agent software, which require machine to take actions to support human for solving issues or dealing with tasks. The most simplify agent was implemented by the rules, however, it is very difficult to input every rule into machine due to the environment complexity, and physical capacity limitations. Algorithms are sometimes implemented to reduce complexity—for example, the least-cost-path algorithm (Collischonn & Pilar, 2000). Since 1970, the application of artificial intelligence in education (AIED) has been a very interesting research topic. Intelligent Tutoring System (ITS) is the most common implementation studied in AIED (du Boulay, 2016). The ITS was studied aiming at identifying at-risk students to monitor the learning behavior of students and generate personalized learning recommendations (Woolf, 2010). ITS has shown considerable improvement in students' performance and outcomes in learning (Ma, Adesope, Nesbit, & Liu, 2014; Schroeder, Adesope, & Gilbert, 2013). In recent years, due to the huge data availability and improved digital technologies, the AIED has been much easier to study and implement. Chen, Xie, and Hwang (2020) systematically aggregated the artificial intelligence-based research performances in the field of education, and the statistical data shows that 74% of the researches was conducted in the past seven years from 1999 to 2019, which indicates the importance of this research topic in recent years. Although the above research summary shows the importance of AIED, Yang (2021) explain that bringing AI into education is not to just apply digital technology into the classroom; educators also need to be human-centric (Yang, 2021). Human-centered artificial intelligence (HAI) is defined as AI under human control and AI on the human condition (Yang, Ogata, Matsui, & Chen, 2021). Especially, the educators also need to be sure that the learners can achieve higher learning performance when AI is reasonably reliable. In practice, Lu, Huang, and Yang (2021) raised a typical case of machines losing reasonable reliability in which machines can ignore some risk students because the teacher adopted a leniency grading policy.

The proposed study was conducted to implement AI-based applications in the education field and focused on practicing the short-answer questions for three reasons. First, Hwang, Xie, Wah, and Gašević (2020) collected the research issues in AIED and found that improving students' learning performance using AI-based solutions is an important research topic that can be second only to designing AI tools. The proposed study believed in short-answer questions, which is one of the best ways to implement AI-based solutions, and the detailed explanation on the implementation is described in the following section. Second, according to prior studies, most research on the ITS has been based on numerical data driven applications, for example, Jovanović, Gašević, Dawson, Pardo and Mirriahi (2017) used students' login times per week as the data to construct the self-regulated learning model. Natural language processing (Chowdhary, 2020), and speech recognition (Deng, Hinton, & Kingsbury, 2013) have been proved to be useful in research, but they have not been adopted in education practice. For our short-answer system, natural language processing is fundamental to its operation. In natural language processing, semantics-based or syntax-based question-generation algorithms can be applied (Greving & Richter, 2018), further discussion of related studies in the Literature Review section. Third, Luckin, Holmes, Griffiths and Forcier (2016) mentioned that most of the current ITS designs are student-oriented. However, the gaps in the current AIED research should also consider the teacher's retention rate on ITS. If teachers are encouraged to design most learning activities on ITS, students can obtain the expected benefits in ITS. Therefore, in the real scenario, if the machine can be used to replace the teacher to evaluate students' learning performance, it is expected to increase the teacher's willingness to use.

## 1.2. The benefits of short-answer questions

The question and answer (Q&A) process yields benefits in many fields; for instance, medical diagnosis and computer system security usefully apply Q&A (Kaur & Bathla, 2015). In the education field, benefits of the Q&A process include (1) allowing student to use question-based practice to construct knowledge, (2) identifying student misunderstandings through learner feedback, (3) guiding learners to pay more attention to key material, (4) repeating concepts to enhance memory, (5) motivating learners to engage in the course, and (6) enabling teachers to understand the learning performance of each learner (Kaur & Bathla, 2015; Kurdi, Leo, Parsia, Sattler, & Al-Emari, 2020). Comparing with other exam methodologies, studies have demonstrated the efficiency of short-answer questions is superior to other modes of examination. For instance, Smith and Karpicke (2014) conducted an experiment with 80 students to investigate the effects of short-answer and multiple-choice questions on retrieval ability, and the results indicated that short-answer questions produced better learning performance due to the higher memory retrieval capability. Rush, Rankin and White (2016) demonstrated that answering short-answer questions requires learners to have a higher level of cognition than answering multiple-choice questions, which means that the learner is required to focus more on the review process. Furthermore, Greving and Richter (2018) recommended short-answer questions because practice with such questions improved student ability to retrieve material from their memories.

Repetitive or frequent requests students to evaluate the knowledge taught in the classroom is an advanced application of short-answer in education, commonly known as "practice testing" or "repeated testing" (Adesope, Trevisan, & Sundararajan, 2017; Wiklund-Hörnqvist, Jonsson, & Nyberg, 2014). Many studies have reported that short-answer-based practice tests elicit strong student performance. For instance, McDermott, Agarwal, D'Antonio, Roediger III, and McDaniel (2014) conducted four experiments with 512 participants and assigned different test plans for the experimental groups; the results indicated that frequent quizzes can improve students' learning outcomes and retention rates. Moreover, Larsen, Butler and Roediger III (2009) investigated the effect of repeated study on final recall, and their experimental results indicated that adopting a repeated study strategy led to higher final scores. Despite its evident value, creating short-answer practice material is time consuming and thus burdensome for teachers. Employing automatic question generation (AQG) may be a solution to this problem; in this technique, question-answer pairs are generated through analysis of a given text (Rus, Cai, & Graesser, 2008), however, the concept that applying AQG techniques into classroom for the practice testing only implemented in the laboratory settings (Greving & Richter, 2018) only.

In summary, implementing AI can benefit students and teachers in the field of education. Due to advancements in its technology, natural language processing may be able to help teachers easily generate short-answer practice tests for classroom use. Therefore, we adopted machine learning to create short-answer questions and investigated whether the generated questions had acceptable quality and whether students benefited from studying with such questions. Our research questions were as follows:

**RQ1:** Can students improve their learning performance with repeated short-answer question practice?

**RQ2:** In evaluating students' programming skill, do machine-generated questions exhibit similar quality to expert-authored questions?

**RQ3:** In evaluating students' programming skill, does machine-grading exhibit similar quality to expert-grading?

## 2. Literature review

AQG can be used to generate various types of questions, such as cloze questions and multiple-choice questions (Ch & Saha, 2018). We used AQG to create short-answer questions because such questions benefit students' long-term memory (Greving & Richter, 2018). The concept of AQG was defined by Le, Kojiri, & Pinkwart (2014) as: "*generating questions from various inputs such as raw text, database or semantic representation*" (p. 352), and it has been a popular research topic in recent years due to the emergence of natural language processing by neural networks, which is designed to mimic how human beings use language and serve as a tool for manipulating human language to meet specific requirements (Chowdhary, 2020). At least three related systematic reviews have been retrieved from the library system in past three years (Ch & Saha, 2018; Kurdi et al., 2020; Papasalouros & Chatzigiannakou, 2018), from which we have gained two valuable information: (1) approaches to implement AQG system, and (2) quality evaluation methods.

According to the systematic review of Kurdi et al. (2020), an AQG system can be implemented through four approaches, but only two of these methods account for more than 70% of instances of implementation. The first one is syntax-based approach, which extracted features such as: part-of-speech, and then select distractors based on a classification algorithm for constructing question sentences (Das & Majumder, 2017). The second approach is based on semantics and depends upon a comprehensive understanding of the context and additional information or knowledge to select meaningful sentences for constructing question sentences (Chan & Fan, 2019; Yao, Bouma, & Zhang, 2012). The other methods are limited by sentence patterns, and therefore, we only considered syntax-based and semantic-based approaches in this study and propose an ensemble method that combines semantic and syntactical approaches to automatically generate questions. For semantics, our system uses BERT (Bidirectional Encoder Representations from Transformers) (Devlin, Chang, Lee, & Toutanova, 2018), the syntax part uses Stanford CoreNLP (Manning, Surdeanu, Bauer, Finkel, Bethard, & McClosky, 2014), and the question construction part uses GPT2 (Generative Pre-Training)(Radford, Narasimhan, Salimans, & Sutskever, 2018). The main reason is that the above methods took transfer learning (Pan & Yang, 2009) approach, which allows follow-up developers to produce a domain specific model without collecting a large dataset, and the methods employed perform well in machine reading comprehension. More details about the collaboration between BERT, Stanford CoreNLP and GPT2 will be introduced in section 0

Another major concern is how to evaluate the quality of AQG methods. Most studies in this field have adopted a standard dataset for evaluating performance, with one of the most popular datasets being the SQuAD (Stanford Question Answering Dataset) (Ch & Saha, 2018), which consists of 100,000 question-answer pairs collected from Wikipedia articles. The SQuAD has been used in BERT, which we used in this study, and several pretraining models such as UNILM (Dong et al., 2019) or Glomo (Yang, Zhao, Dhingra, He, Cohen, William, Salakhutdinov & LeCun, 2018). On the other hand, how did priori studies quantify the performance evaluation results? Practically, a comparison will be performed between the questions generated through the proposed method and some other methods from related works, and the questions in SQuAD dataset will be served as the ground truth during the comparison process. Various metrics will be used for the quantify the comparison results—for example, BLEU (Bilingual Evaluation Understudy) (Papineni, Roukos, Ward, & Zhu, 2002) or ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004).

Through the above studies, we can observe that BERT, CoreNLP and GPT2 exhibit outstanding performance in semantic-based and syntax-based question generation (Klein & Nabi, 2019). The results guide us to consider those approaches as the implementation of the AQG methods. However, because the input to those pre-training models was the teacher's teaching material, no standard answers or ground truth were available for BLEU or ROUGE to use to evaluate the quality of interrogative sentences. Therefore, we reviewed the most typical evaluation approach: the Turing test (Turing, 2009), which was proposed by Alan Turing in 1950. In the test process, Turing suggest to assign an evaluator to judge the messages that delivered by human or machine through a dialog, and expect the evaluator cannot judge the difference between human and machine due to the machine present similar response as human. Several studies adopted Turing test in order to prove the performance; for example, (Hingston, 2009) designed a game bot, and after five rounds of games where machines imitated humans, they analyzed game behavior. The result of the analysis declared: "*Computers cannot play like humans—yet.*" In another instance, Alarifi, Alsaleh, & Al-Salman (2016) proposed a classifier by using the graph theory to detect fake identities on social network. To demonstrate the performance under the situation that lack of the ground truth datasets, they used the Turing test as well. Finally, in the summary report compiled by Kurdi et al. (2020), they also recommended using the expert review process for assessing machine-generated

questions. Thus, to evaluate the quality of our AQG method, we adopted the Turing test approach, which is detailed in the next section.

### 3. Method and experiments

#### 3.1. Participants

This study conducted an empirical experiment to assess the impact of practice testing on learning performance. The experiment was executed in a freshman university programming course during three weeks in October 2020, and the primary course content was the basic Python programming language. A total of 91 students from two classes participated in the experiment. All the participants were students in the computer science department. We divided the students into two groups: the first one was the control group with 50 students, and they learned through conventional learning activities; the second one was the experimental group with 41 students, and they learned through practice testing.

#### 3.2. Experimental design and learning activities

The design of the experiment adopted in this study is shown in Figure 1. The learning activity was divided into three phases: initial phase, course instruction, and performance evaluation. Before the course began, the teacher uploaded the learning materials to **BookRoll** (Flanagan & Ogata, 2017). In the **initial phase**, the teacher assigned a pre-test to evaluate the students' programming skills at the beginning of the course; in the next step, instruction in the if-else programming syntax was given and practice was assigned to the experimental group. The practice in this step was delivered by the short-answer system we proposed in this study (explained further in the next section). The second phase was the **course instruction phase**, during which only general classroom activities were conducted and the experimental group completed the practice test generated by the machine, as in the initial phase. The third phase was the **performance evaluation phase**. In addition to regular instruction activities and practice tests for the experimental group, both groups took a post-test to evaluate their programming skill; the short-answer questions used in the post-test are listed in the Appendix I. The grading results of the pre-test and post-test were compared. Furthermore, the post-test was graded by both experts and machine, respectively, and the grading results by expert will be the ground truth to (1) compare with the pre-test to investigate students' learning performance improvements and (2) compare with machine-grading to evaluate the quality of it.

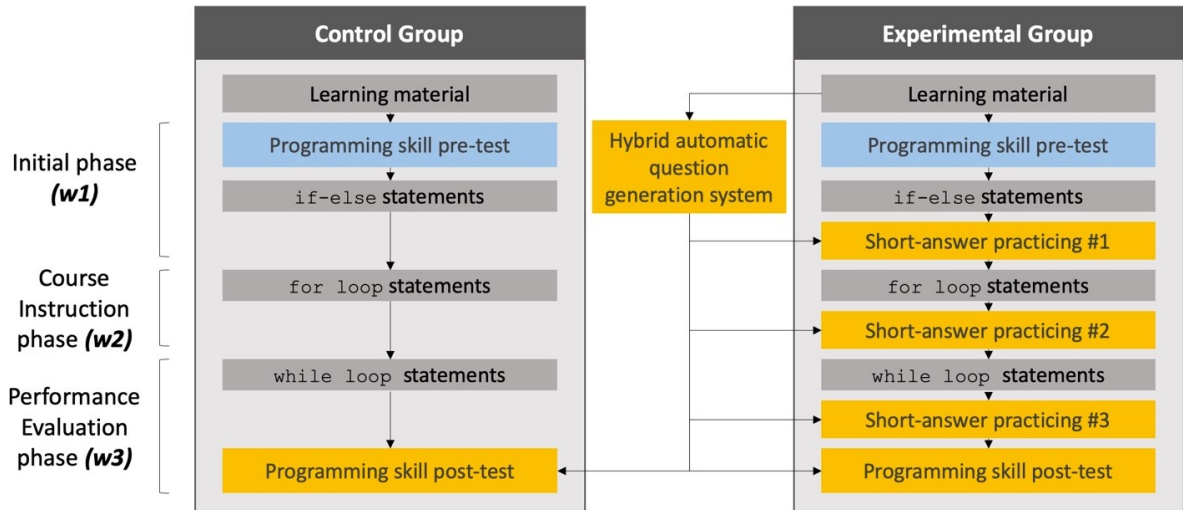


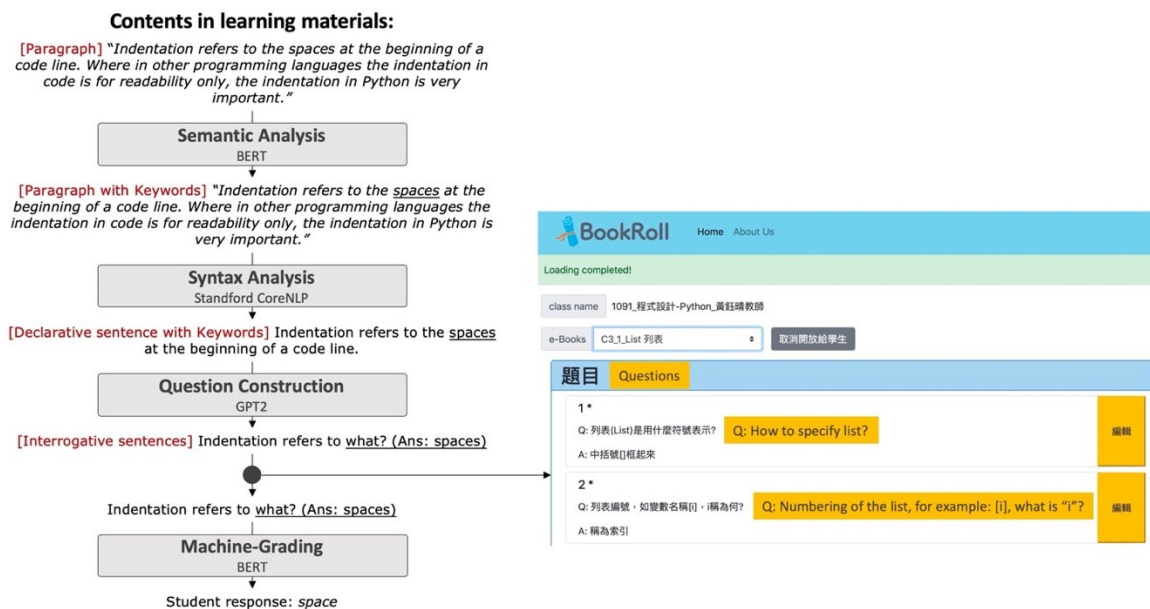
Figure 1. Experimental design and learning activities

#### 3.3. Hybrid automatic question generation (Hybrid-AQG) system

The AQG system proposed by this study was combined semantic analysis and syntax analysis, it provides two functions: machine-questioning and machine-grading. Figure 2 shows the user interface designed in this study, which allows the instructor to review and modify questions generated by the machine and students to respond to the questions. The design principle of machine questioning is to let the machine understand the learning

material’s content and generates question sentences. It consists of semantic analysis and syntactic analysis; therefore, the system here we named as Hybrid-AQG, and we listed questions that generated by the system in the Appendix II.

Because of the syntactical analysis output a declarative sentence with keywords only, but we only need a sentence without keywords to generate the interrogative sentences. Therefore, we transformed the declarative sentences into interrogative sentences. Practically, we first removed the keyword specified by semantic analysis, and then we adopted the GPT2 (Radford et al., 2018) which is a machine learning technology that uses unsupervised learning to generate reasonable words according to the meaning of the context. As shown in Figure 2, we fed a sentence into the GPT2 model, and the model predicted the following word, “what”; thus, an interrogative sentence was achieved.



### 3.4. Evaluation the quality of machine-question and machine-grading



thus, the same expert scored both the pre-test and post-test. The following two subsections and Figure 3 explain the details of the assessment of items two and three.

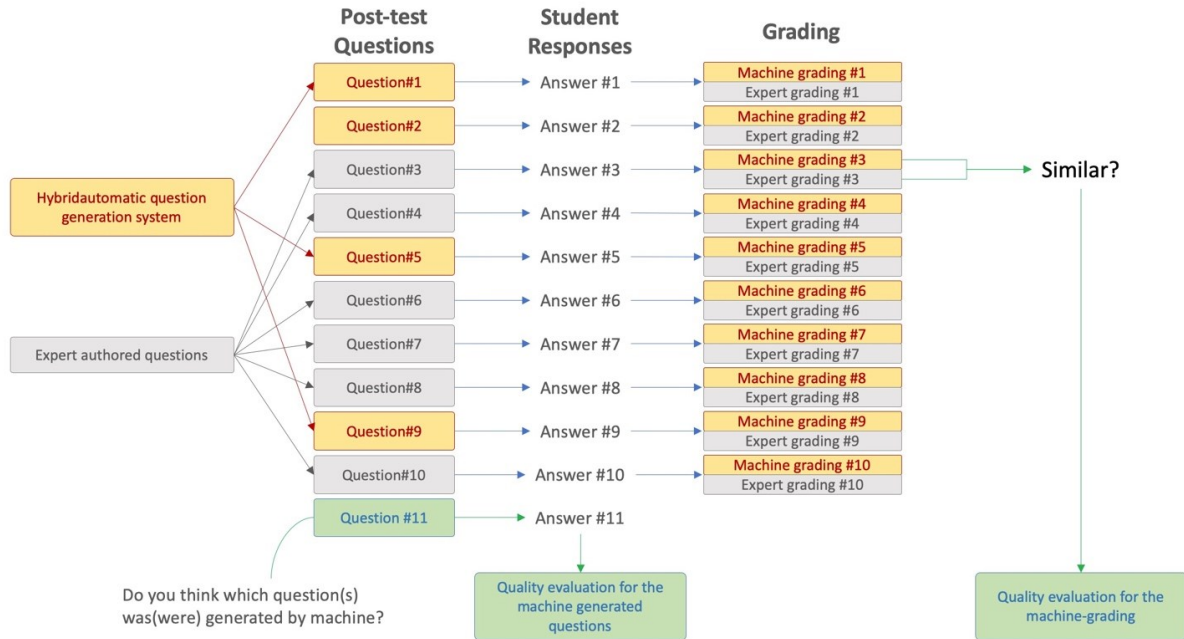


Figure 3. Flow in post-test for evaluating the quality of the machine-generated questions and machine-grading

Inspired by the Turing test, to evaluate the quality of machine-generated questions, we designed a test that featured both machine-generated questions and expert-authored questions and then evaluated whether students could distinguish them. To evaluate students' programming skills in the post-test, a total of 11 questions were presented; except for Questions 1, 2, 5, and 9, the rest were expert-authored questions. The main reason for only using machine question generation for 4 of 10 questions was because the Turing test requires machines to be involved in just 1/3 to 1/4 of an entire test. Further, in Question 11, we asked students to identify which question(s) was(were) generated by a machine to determine whether they could distinguish authorship. By contrast, if the machine-generated questions are qualified, we expected that the students could not answer the correct answer. Finally, in the post-test stage, to correctly quantify the students' programming skills and compare their results with their pre-test scores, the results of Question 11 were not considered.

Inspired by the Turing test, to evaluate the quality of machine-generated questions, we designed a test that featured both machine-generated questions and expert-authored questions and then evaluated whether students could distinguish them. To evaluate students' programming skills in the post-test, a total of 11 questions were presented; except for Questions 1, 2, 5, and 9, the rest were expert-authored questions. The main reason for only using machine question generation for 4 of 10 questions was because the Turing test requires machines to be involved in just 1/3 to 1/4 of an entire test. Further, in Question 11, we asked students to identify which question(s) was(were) generated by a machine to determine whether they could distinguish authorship. By contrast, if the machine-generated questions are qualified, we expected that the students could not answer the correct answer. Finally, in the post-test stage, to correctly quantify the students' programming skills and compare their results with their pre-test scores, the results of Question 11 were not considered.

Table 1. Confusion matrix to evaluate the quality of machine-questioning and machine-grading

	Evaluating machine-questioning quality		Evaluating machine-grading quality	
	Student distinguish	Actual	Machine classified	Expert confirmed
True-Positive (TP)	Machine-generated	Machine-generated	Correct	Correct
False-Positive (FP)	Machine-generated	Expert-authored	Correct	Incorrect
False-Negative (FN)	Expert-authored	Machine-generated	Incorrect	Correct
True-Negative (TN)	Expert-authored	Expert-authored	Incorrect	Incorrect

To quantify machine-generated question quality, we treated the answers to Question 11 as a binary classification problem and applied a confusion matrix for comparison. Four combinations are listed in Table 1, and we calculated accuracy, precision, and recall in light of responses designated as true positive (TP), false positive

(FP), true negative (TN), or false negative (FN) by using the following equations. This test used accuracy, recall and precision to evaluate the quality of machine-generated questions.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

In evaluating machine-generated question quality by using the confusion matrix, accuracy indicated the ratio of the 10 questions correctly identified. If the accuracy was close to 1, the students could distinguish the questions generated by the machine, which would mean machine performance was not acceptable. However, if accuracy was close to 0.5, this would mean the quality of machine-generated questions approached that of expert-authored questions. Recall indicated the proportion of correctly identified machine-generated questions. A higher recall value indicated a higher rate of correctly identified questions. Precision referred to the proportion of number of questions that students think is machine-generated in actual number of machine-generated questions. The higher the precision value was, the higher was the ratio of correctly identified machine-generated questions.

To evaluate whether machine and expert grading was of similar quality, we adopted a confusion matrix, too. For each answer, whether an expert or a machine checked it, a binary result was given: “correct” or “incorrect.” Then, we took the confirmed results from the expert as the ground truth; the four combinations are listed in Table 1. Accuracy, recall, and precision were again applied for assessment. Accuracy close to 1 suggested close similarity between expert and machine grading, but accuracy close to 0.5 suggested inconsistency. Recall indicated the ratio of answers correctly graded by experts to those correctly graded by machine learning. Precision indicated the ratio of correct answers verified by expert-grading.

## 4. Results

### 4.1. Reply RQ1 (Can students improve their learning performance with repeated short-answer question practicing?)

To measure students’ learning performance, the teacher conducted a pre-test and post-test to evaluate programming skills in the first week and the third week. We used the independent samples *t*-test to assess the difference in learning performance between the control and experimental groups in the pre-test. Table 2 lists the results of the descriptive statistics and independent samples *t*-test of the pre-tests of the control and experimental groups. The scores for the pre-test of programming skills in the control and experimental groups were 77.0 and 73.38, respectively. The results listed in Table 2 indicate that the pre-test scores for the experimental group and control group did not differ significantly ( $t = -1.117$ ,  $p > .05$ ). This means that the students’ programming skills were equal in the control and experimental groups.

Table 2. Statistics results and independent sample *t*-test of pre-test for the control group and the experimental group

Group	<i>N</i>	Mean	<i>S.D.</i>	<i>t</i>	<i>p</i>
Experimental group	41	73.38	16.76	-1.117	.267
Control group	50	77.0	14.18		

Note. \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ .

This study investigated the impact of repetitive short-answer practice on students’ learning performance by using analysis of covariance (ANCOVA) to exclude the difference in programming skills of the control and experimental groups. The pre-test and post-test scores for programming skills were used as the covariate and dependent variables in ANCOVA, respectively. The result of Levene’s test did not violate the homogeneity of variance ( $F = 601$ ,  $p = .440$ ), meaning that ANCOVA was applicable.

Table 3. Statistics results and ANCOVA of post-test for the control group and the experimental group

Group	<i>N</i>	Mean	<i>S.D.</i>	Adjusted Mean	<i>S.D. Error</i>	<i>F</i>	<i>p</i>
Experimental group	41	88.78	10.97	89.12	2.21	12.73	.000***
Control group	50	78.75	16.42	78.74	2.00		

Note. \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ .

Table 3 presents the descriptive statistics results and ANCOVA of the post-test for the control group and the experimental group. The adjusted means of the post-test scores in programming skills for the control and experimental groups were 78.74 and 89.12, respectively. According to the ANCOVA result, the experimental group had significantly higher post-test scores than did the control group ( $F = 12.73$ ,  $p = .00$ ). The results demonstrated that students can effectively improve their learning performance in programming skills through use of the Hybrid-AQG, or more specifically, the repetitive short-answer practice by machine-generated questions. Our results were consistent with those of prior studies that found that repetitive practice can enhance students' long-term memory to drive subsequent improvements in learning performance (Karpicke, 2017; Roediger III & Karpicke, 2006; Rowland, 2014), especially when short-answer practice is applied in the higher education context (Greving & Richter, 2018). Moreover, it is inevitable for students to be familiar with the topic for their performance, but it does not mean that the content of the questions is qualified. Therefore, we will continue to discuss the quality of machine-questioning and machine-grading in the following sessions.

#### 4.2. Reply RQ2 (To evaluating students' programming skill, does machine-generated questions have similar quality with the expert-authored questions?)

This study adopted an evaluation process based on the Turing test to investigate the ability of students to identify machine-generated questions. The teacher designed Question 11 in the post-test, which asked students to identify which questions were generated by a machine. Figure 4 presents the results for the ability of students to distinguish machine- from expert-authored questions. Four questions, namely 1, 2, 5, and 9, generated by machine were correctly distinguished by 13 (32%), 22 (54%), 12 (29%), and 12 (29%) students, respectively, in the experimental group, and 9 (18%), 12 (24%), 9 (18%), and 16 (32%), respectively, in the control group. These results indicate that in experimental group, a higher proportion of students could correctly distinguish between the machine- and expert-authored questions, which we attribute to the students in the experimental group having already seen the patterns of machine-generated questions when using the short-answer practice system. This provides evidence that the experimental group students created long-term memories during repetitive practice.

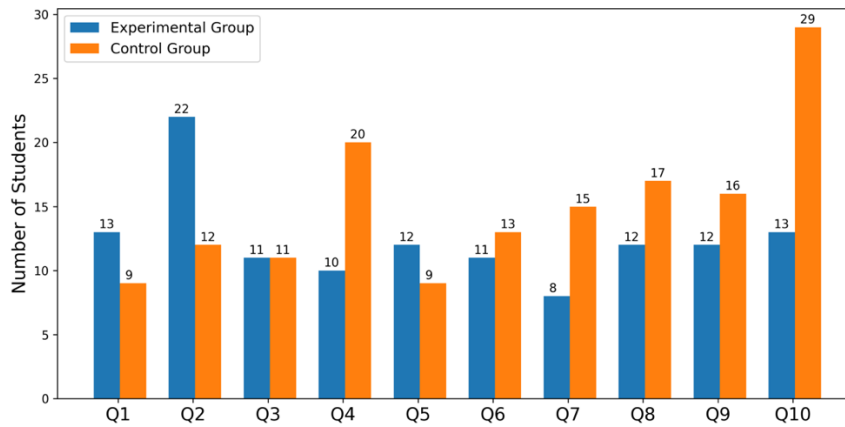


Figure 4. Distinguishing student results for machine- and expert-authored questions

We applied the confusion matrix to quantify the performance of students in distinguishing between machine-generated and expert-authored questions in the post-test to evaluate the quality of the machine-generated questions. Figure 5 presents the accuracy, precision, and recall of guessing results for the experimental and control groups in the post-test. The accuracy ( $t = -3.7$ ,  $p < .001$ ), precision ( $t = -2.48$ ,  $p < .05$ ), and recall ( $t = -2.53$ ,  $p < .05$ ) values of the experimental group were significantly higher than those of the control group.

The average accuracy of the control group is .48, which means that the control group answered questions 11 almost answering by guessing. Thus, the students in the control group could not distinguish which questions were generated by a machine. Compared with the control group, the experimental group exhibited a higher average accuracy: .585. Studies have indicated that practice can enable students to construct knowledge and that repeat practice using short-answer questions can enhance students' retrieval of information from memory (Kaur

& Bathla, 2015; Kurdi et al., 2020). We attribute the higher values of accuracy, recall, and precision in the experimental group to the students in the experimental group having had practice with similar machine-generated questions in the Hybrid-AQG system; such practice had a positive effect on their quality of review and deepened their long-term memory of the machine-generated questions. This is consistent with the observation in Greiving and Richter (2018) study that short-answer practice can help students retrieve material from memory. However, even though the students in the experimental group had seen the machine-generated questions, the accuracy, precision, and recall still only reached .585, .436, and .36, respectively, indicating that the machine-generated questions were similar to expert-authored questions. Thus, we conclude that students perceive machine-generated questions and expert-authored questions similarly, indicating that the machine-generated questions are suitable for practice testing.

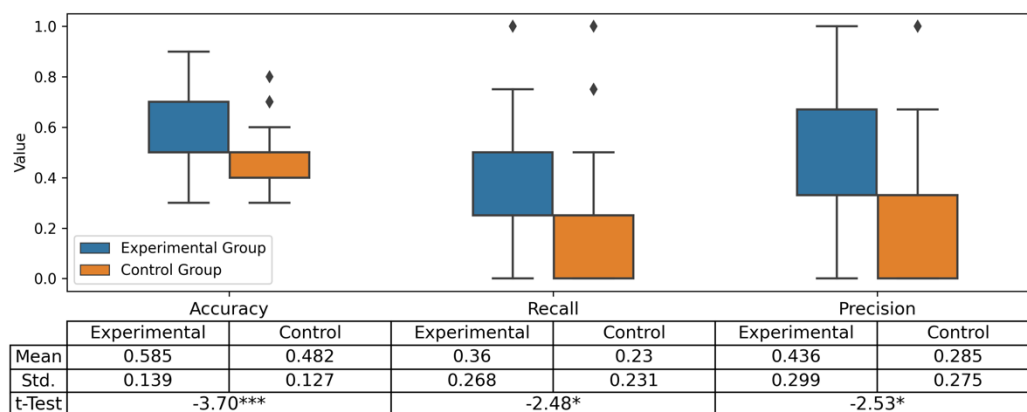


Figure 5. Distinguishing experimental and control group results for machine- and expert-authored questions by using the metrics of accuracy, precision, and recall

Recall was evaluated as the ratio of guesses that correctly identified the four machine-generated questions. For students in the experimental group, the values of TP, FP, FN, and TN were 59, 65, 105, and 181, respectively. For students in the control group, the values of TP, FP, FN, and TN were 46, 105, 154, and 195, respectively. Figure 5 indicates that the precision values for the experimental group and the control group (experimental group: .436; control group: .285) were higher than the recall values (experimental group: .36; control group: 0.23), meaning that students in both the experimental group and the control group struggled to distinguish which questions were machine-generated. In both the experimental and control groups, the values for precision were higher than those for recall, as evidenced by fewer FPs than FNs. FN meant that a student distinguished that the question was generated by an expert, but it was actually a machine-generated question. Higher FN values indicated that students tended to treat machine-generated questions as expert-authored questions. This may have been because machine-generated questions were similar to expert-authored questions, and thus, students struggled to distinguish them—hence, the lower recall value. This outcome indicates that the text content used by the machine when generating the questions (using teaching materials and natural language processing) was quite close to the text content used by the expert when designing questions, meaning that the machine-generated questions in this study are suitable for practice testing due to students being unable to distinguish between machine-generated and expert-authored questions.

Table 4. Correlation analysis of the number of correct answers and the number of students who identified the question as machine generated

	Mean/Std. of students		Spearman correlation	
	Number of answer the question correctly	Number of students identified machine-generated question	Coefficient	p-value
Experimental	35.50/4.53	9.60/3.84	.09	.79
Control	30.80/12.88	9.90/7.81	.83	.003**

Note. \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ .

To explore why the students are hard to distinguish the questions are generated from machine or expert, the teacher interviewed the control group students how they setup the identification rules. Most students replied that computers are not as smart as humans, and therefore, they adopted an identification rule that sought the simplest questions in the list. Therefore, in order to continuous explore the quality of machine-questioning, we have to proof students in the control group adopted an identification rule like looking for the simplest question, we use Spearman correlation analysis to explore the relationship between the number of students answering correctly and the number of students who identified that the question is belongs to machine-generated. Table 4 lists the

descriptive statistics of Spearman correlation analysis results between answering correlation analysis between the rate of answer the question correctly and machine-expert identification rate.

As evident in Table 4, the number of students in the control group who answered a question correctly and the number of students who identified the question as being machine generated had a significant correlation ( $r = .83$ ,  $p > .01$ ). This result shows that for the students in the control group, more students answered simple questions correctly, which they identified as machine generated. This finding is consistent with what the students described as their identification rule. This suggests that without the benefit of the Hybrid-AQG system, the students defaulted to identifying machine-generated questions by their simplicity. By contrast, the experimental group exhibited no such correlation between correct answers and machine-generated question identification ( $r = .09$ ,  $p > .05$ ), which we attribute to students having practiced with the Hybrid-AQG system, enabling students to identify whether a question was machine-generated or expert-authored based on memory. This phenomenon evident in the experimental group is consistent with research results (Greving & Richter, 2018) indicating that the Hybrid-AQG system can enable students to retrieve more material from memory.

#### 4.3. Reply RQ3 (To evaluating students' programming skill, does machine-grading have the similar quality with the expert-grading?)

To measure the quality of machine grading in the post-test, this study used a confusion matrix to evaluate the difference between machine grading and expert grading. The process of evaluating machine-generated and expert-authored questions was described in detail in Methods section. Figure 6 presents the accuracy, recall, and precision of machine grading quality; the accuracy ( $t = 4.135$ ,  $p < .001$ ), precision ( $t = 2.084$ ,  $p < .05$ ), and recall ( $t = 4.689$ ,  $p < .001$ ) values for the experimental group were significantly higher than those for the control group.

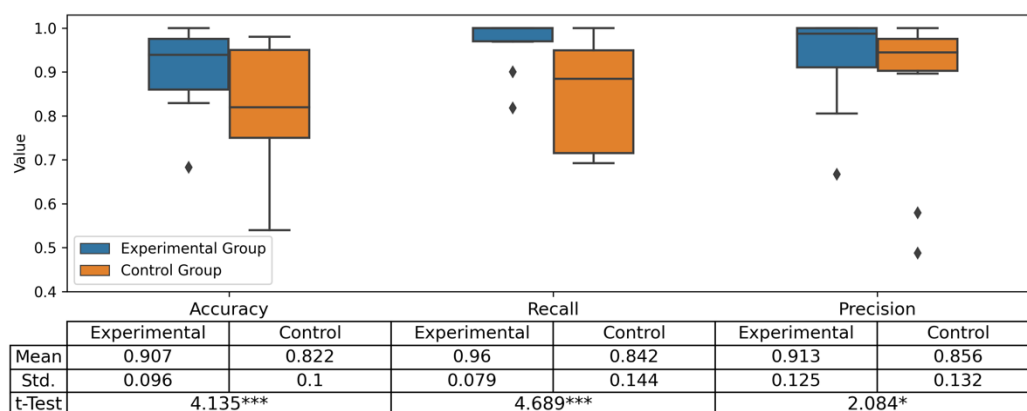


Figure 6. Similarity analysis between expert-grading and machine-grading, in metrics of accuracy, precision and recall in between experimental and control groups

The accuracy of the experimental group and the control group was .907 and .822, respectively. The results of the independent samples  $t$ -test of accuracy demonstrate that the experimental group was significantly more accurate than the control group ( $t = 4.135$ ,  $p < .001$ ). The main reason is that the content answered by the students in the experimental group makes the machine more interpretable. This may be the result of repeated practice by the students in the classroom. We infer that repeated practice using the Hybrid-AQG system can enhance the long-term memory of students, which is why the accuracy of machine grading and expert grading of the experimental group was higher than that for the students in the control group. This result combined with the results of the analysis of RQ1, which indicated that students in the experimental group had significantly higher learning performance in the post-test than students in control group did, lead us to conclude that practicing short-answer questions can enhance the long-term memory of students (as evidenced by the experimental group performance) and further improve their academic performance. These benefits of the Hybrid-AQG system are consistent with the results of (Greving & Richter, 2018), which suggested that repetitive practice can enhance student's ability to retrieve information from memory.

In this study, the recall values of the experimental group and the control group were .96 and .84, respectively, meaning that machine grading and expert grading were highly consistent for correct answers; thus, machine grading can replace expert grading to some extent. In this study, the precision values of the experimental group and the control group were 0.91 and .85, respectively, meaning that machine and expert grading are highly consistent for answers that a machine grades as correct.



For the experimental group, the values of TP, FP, FN, and TN were 318, 27, 11 and 54, respectively. The value of recall was higher than that of precision due to fewer FNs than FPs. FPs may have been because the machine identified the correct answer, but the answer contained only conceptual keywords rather than complete and clear content, causing experts to think the answer was wrong. To investigate the reasons for the FP-type answer, we examined students' FP-type answers and found that because the answer content contained keywords related to the concepts covered by the question, it deemed the correct answer in machine grading; however, because the answer content was not complete, experts graded it as a wrong answer. For example, one question asked, "*What is on the right side of the equal sign when assigning a value to a variable?*" In the FP answer, the content example was the actual content value. The machine-based grading was mainly based on whether the content value was mentioned in the answer, and thus, the machine-grading system rated this answer as correct, but the expert thought that the content value of the variable must be clearly stated, and thus, the answer was considered to be incomplete and rated as wrong. From this example, we suggest that the expert grading may be stricter than the machine-based grading. This may account for why the number of FPs was greater the number of FNs and may also be the reason that the value of recall was greater than precision in the experimental group.

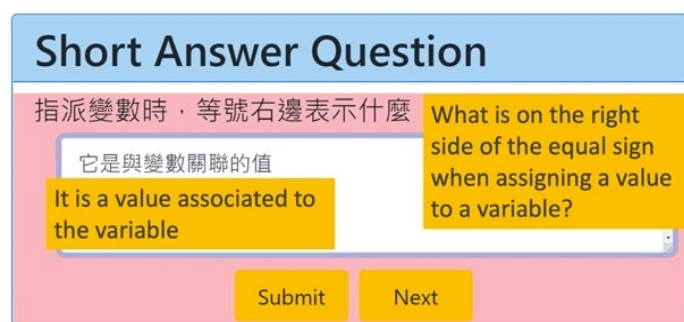


Figure 7. False-Negative machine grading example

For the students in the control group, the values of TP, FP, FN, and TN were 257, 38, 51, and 154, respectively. Because fewer FPs were recorded than FNs, the precision value was higher than the recall value. FN indicated the machine classified the answer as incorrect, but the expert confirmed the answer as correct. Here we provide an example as shown in Figure 7: "*What is on the right side of the equal sign when assigning a value to a variable?*" One FN answer from student is: "*It is a value associated to the variable,*" and the issue in this sentence is the pronoun "it". Because experts know that the pronoun "it" in the answer refers to a variable but the machine learning system has no context to conclude this, it fails to understand what the pronoun in the answer refers to. Therefore, the machine-based grading regards the answer as wrong. For the students in the control group, because they did not use the short-answer practice system and had never experienced machine grading, they did not know how to answer the answer with content that the machine could understand. Therefore, most answers were incorrectly scored in machine grading. This is why the precision value of the students in the control group was higher than their recall value.

## 5. Conclusions

The technical applications of modern AI are diverse, such as computer vision or speech recognition. This study focuses on natural language processing, aims to implement AI applications for the education purpose and look forward to the benefits of emerging AI technologies that can bring into education. To this end, this study proposed Hybrid-AQG based on the advanced transfer learning technologies BERT, GPT2, and CoreNLP. The system can perform semantic and syntactical analysis of a teacher's teaching materials, generate multiple question-answer pairs, and enables students to engage in repeat practice of questions after class. Through implementation of this system, a teacher's burden is reduced and students' long-term memory of course content can be enhanced.

After a 3-week experiment, we verified three hypotheses through data analysis. First, repetitive practice was proven to be beneficial to students' long-term memory for subsequent improvements in learning performance, even when using practice questions generated by a machine. Second, in our experiment, only some students could identify when a question was machine rather than expert generated because of long-term memory; however, most students could not distinguish them. This reflects the maturity and usefulness of combining semantic and syntax approaches for generating questions. In the control group, students simply defaulted to identifying simple questions a machine generated. Last, this study employed a semantic method to implement the machine-grading functionality. However, it turned out that grading short-answer questions still requires

technology that can understand the context and order of keywords, otherwise, only keywords check can be achieved in current study.

Finally, there are two limitations to this study. The first is that we didn't discuss the teaching materials provided by teachers. Still, the quality of the teaching materials and the format setting will affect the machine-generated questions' quality. For example, some teachers preferred to use pictures and even sample code in teaching materials, and both ways presentation will cause some garbled information in the output of machine-questioning, and it required the teacher to review and remove. The second limitation is the issue of the question-type. In this study, we only used short-answer questions; however, other popular types need to be verified the effectiveness and quality, such as the multiple-choice questions and cloze questions.

## Acknowledgement

This work is supported by Ministry of Science and Technology, Taiwan under grants MOST-109-2511-H-008-007-MY3, MOST-108-2511-H-008-009-MY3, MOST-110-2511-H-153-001, and Ministry of Education, Taiwan.

## References

- Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A Meta-analysis of practice testing. *Review of Educational Research*, 87(3), 659–701.
- Alarifi, A., Alsaleh, M., & Al-Salman, A. (2016). Twitter turing test: Identifying social machines. *Information Sciences*, 372, 332–346.
- Ch, D. R., & Saha, S. K. (2018). Automatic multiple choice question generation from text: A Survey. *IEEE Transactions on Learning Technologies*, 13(1), 14–25.
- Chan, Y.-H., & Fan, Y.-C. (2019). A Recurrent BERT-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering* (pp. 154–162). doi:10.18653/v1/D19-5821
- Chen, X., Xie, H., & Hwang, G.-J. (2020). A Multi-perspective study on artificial intelligence in education: Grants, conferences, journals, software tools, institutions, and researchers. *Computers and Education: Artificial Intelligence*, 100005. doi:10.1016/j.caeai.2020.100005
- Chowdhary, K. R. (2020). Natural language processing. In *Fundamentals of artificial intelligence* (pp. 603–649). Springer. doi:10.1007/978-81-322-3972-7\_19
- Collischonn, W., & Pilar, J. V. (2000). A Direction dependent least-cost-path algorithm for roads and canals. *International Journal of Geographical Information Science*, 14(4), 397–406.
- Das, B., & Majumder, M. (2017). Factual open cloze question generation for assessment of learner's knowledge. *International Journal of Educational Technology in Higher Education*, 14(1), 1–12.
- Deng, L., Hinton, G., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An Overview. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8599–8603). doi:10.1109/ICASSP.2013.6639344
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. Retrieved from <https://arxiv.org/abs/1810.04805>
- Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., & Hon, H.-W. (2019). Unified language model pre-training for natural language understanding and generation. Retrieved from <https://arxiv.org/abs/1905.03197>
- du Boulay, B. (2016). Artificial intelligence as an effective classroom assistant. *IEEE Intelligent Systems*, 31(6), 76–81.
- Flanagan, B., & Ogata, H. (2017). Integration of learning analytics research and production systems while protecting privacy. In *Proceedings of the 25th International Conference on Computers in Education* (pp. 333–338). Christchurch, New Zealand: Asia-Pacific Society for Computers in Education.
- Greving, S., & Richter, T. (2018). Examining the testing effect in university teaching: Retrieval and question format matter. *Frontiers in Psychology*, 9, 2412. doi:10.3389/fpsyg.2018.02412
- Hingston, P. (2009). A Turing test for computer game bots. *IEEE Transactions on Computational Intelligence and AI in Games*, 1(3), 169–186.

- Hwang, G.-J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 1, 100001. doi:10.1016/j.caeai.2020.100001
- Jovanović, J., Gašević, D., Dawson, S., Pardo, A., & Mirriahi, N. (2017). Learning analytics to unveil learning strategies in a flipped classroom. *The Internet and Higher Education*, 33(4), 74–85.
- Karpicke, J. D. (2017). Retrieval-based learning: A Decade of progress. In *Learning and Memory: A Comprehensive Reference* (2nd ed., pp. 487-514), Cambridge, MA: Elsevier.
- Kaur, J., & Bathla, A. K. (2015). A Review on automatic question generation system from a given Hindi text. *International Journal of Research in Computer Applications and Robotics (IJRCAR)*, 3(6), 87–92.
- Klein, T., & Nabi, M. (2019). Learning to answer by learning to ask: Getting the best of GPT-2 and Bert worlds. Retrieved from <https://arxiv.org/abs/1911.02365>
- Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A Systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1), 121–204.
- Larsen, D. P., Butler, A. C., & Roediger III, H. L. (2009). Repeated testing improves long-term retention relative to repeated study: A Randomised controlled trial. *Medical Education*, 43(12), 1174–1181.
- Le, N. T., Kojiri, T., & Pinkwart, N. (2014). Automatic question generation for educational applications—the state of art. In *Advanced computational methods for knowledge engineering* (pp. 325-338). doi:10.1007/978-3-319-06569-4\_24
- Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop* (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics.
- Lu, O. H., Huang, A. Y., & Yang, S. J. (2021). Impact of teachers' grading policy on the identification of at-risk students in learning analytics. *Computers & Education*, 163, 104109. doi:10.1016/j.compedu.2020.104109
- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An Argument for AI in education*. London, UK: Pearson Education.
- 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.
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). Baltimore, MD: Association for Computational Linguistics.
- McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger III, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, 20(1), 3-21.
- Pan, S. J., & Yang, Q. (2009). A Survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345–1359.
- Papasalouros, A., & Chatzigiannakou, M. (2018, July). *Semantic web and question generation: An Overview of the state of the art*. Paper presented at the International Association for Development of the Information Society (IADIS) International Conference on e-Learning, Madrid, Spain.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A Method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Retrieved from <https://www.aclweb.org/anthology/P02-1040.pdf>
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. Retrieved from <https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf>
- Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255.
- Rowland, C. A. (2014). The Effect of testing versus restudy on retention: A Meta-analytic review of the testing effect. *Psychological Bulletin*, 140(6), 1432- 1463. doi:10.1037/a0037559
- Rus, V., Cai, Z., & Graesser, A. (2008). Question generation: Example of a multi-year evaluation campaign. In *Proceedings of the WS on the QGSTEC*. Retrieved from <https://www.cs.memphis.edu/~vrus/questiongeneration/5-RusEtAl-QG08.pdf>
- Rush, B. R., Rankin, D. C., & White, B. J. (2016). The Impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education*, 16(1), 1–10.
- Russell, S., & Norvig, P. (2002). Artificial intelligence: A Modern approach [PowerPoint slides]. Retrieved from <https://storage.googleapis.com/pub-tools-public-publication-data/pdf/27702.pdf>



- Schroeder, N. L., Adesope, O. O., & Gilbert, R. B. (2013). How effective are pedagogical agents for learning? A Meta-analytic review. *Journal of Educational Computing Research*, 49(1), 1–39.
- Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory*, 22(7), 784–802.
- Turing, A. M. (2009). Computing machinery and intelligence. In *Parsing the Turing test* (pp. 23–65). Springer. doi:10.1007/978-1-4020-6710-5\_3
- Wiklund-Hörnqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology*, 55(1), 10–16.
- Woolf, B. P. (2010). *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Burlington, MA: Morgan Kaufmann.
- Yang, S. J. H. (2021). Guest Editorial: Precision education - A New challenge for AI in education. *Educational Technology & Society*, 24(1), 105-108.
- Yang, S. J. H., Ogata, H., Matsui, T., & Chen, N.-S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 2, 100008. doi:10.1016/j.caeai.2021.100008
- Yang, Z., Zhao, J., Dhingra, B., He, K., Cohen, W. W., Salakhutdinov, R., & LeCun, Y. (2018). Glomo: Unsupervisedly learned relational graphs as transferable representations. Retrieved from <https://arxiv.org/abs/1806.05662>
- Yao, X., Bouma, G., & Zhang, Y. (2012). Semantics-based question generation and implementation. *Dialogue & Discourse*, 3(2), 11–42.

## Appendix I: Post-test questions (where \* means the question was generated by machine)

1. \*What are the rules of naming variables in Python?
2. \*What is the left side of equal symbol when assigning a variable?
3. What is “/” and “\*” means?
4. What is the default value of sep parameter in print() function?
5. \*What is the data type after a division operation?
6. How to present a Python list in symbol?
7. What is the len() function for a Python list?
8. The program would execute “if” or “else” if the condition is not True?
9. \*What is the order of the and 、 or 、 not operator?
10. What is the for loop means?

## Appendix II: Machine-generated questions (where \* means the question also listed in the post-test)

1. What format will the program source code be saved in?
2. What does the computer hope to be able to do?
3. What is wrong with the program syntax?
4. What does the interpreter do?
5. What is wrong with the program logic?
6. What is the purpose of learning languages?
7. What are the grammatical errors?
8. \*What is the left side of equal symbol when assigning a variable?
9. What is the right side of equal symbol when assigning a variable?
10. What is the variable to set the content value through?
11. \*What are the rules of naming variables in Python?
12. What is the variable name?
13. What does String not support?
14. What is stored in memory?
15. What is the data enclosed in?
16. What does the program print the result through?
17. What are the two types of arithmetic operators?
18. What is the highest priority in arithmetic operators?
19. \*What is the data type after a division operation?
20. What is the purpose of using single or double quotes?
21. What is the purpose of using continue in the loop?
22. What is the difference between for loop and while loop?
23. What are the definitions of regional variables and global variables?
24. What can happen to the content in the list?
25. What does List use to hold elements?
26. \*What is the order of the and 、 or 、 not operator?
27. What is the relationship between the grid and the number in the list?
28. How to set if conditional?
29. What does the string use to specify specific characters?
30. What is the string made of?

## Effects of Personalized Intervention on Collaborative Knowledge Building, Group Performance, Socially Shared Metacognitive Regulation, and Cognitive Load in Computer-Supported Collaborative Learning

Lanqin Zheng\*, Lu Zhong, Jiayu Niu, Miaolang Long and Jiayi Zhao

School of Educational Technology, Faculty of Education, Beijing Normal University, Beijing, China //

bnuzhenglq@bnu.edu.cn // bnuzhongl@mail.bnu.edu.cn // 201921010201@mail.bnu.edu.cn //

202021010199@mail.bnu.edu.cn // jiayizhao@mail.bnu.edu.cn

\*Corresponding author

**ABSTRACT:** In recent years, the rapid development of artificial intelligence has increased the power of personalized learning. This study aimed to provide personalized intervention for each group participating in computer-supported collaborative learning. The personalized intervention adopted a deep neural network model, Bidirectional Encoder Representations from Transformers (BERT), to automatically classify online discussion transcripts and provide classification results in real time. Personalized feedback and recommendations were automatically generated from the classification results. A quasi-experimental research design was adopted to examine the effects of the proposed personalized intervention approach on collaborative knowledge building, group performance, socially shared metacognitive regulation, and cognitive load. Sixty-six college students participated in this study and were randomly assigned to the experimental and control groups. For online collaborative learning, students in the experimental group adopted the personalized intervention approach, whereas those in the control group used the conventional approach. Both quantitative and qualitative research methods were adopted to analyze data. The results indicated significant differences in the level of collaborative knowledge building and group performance between the experimental and control groups. Furthermore, the experimental group demonstrated more socially shared metacognitive regulation than the control group. There was no significant difference in cognitive load between the experimental and control groups. The results obtained from interviews were consistent with the quantitative data. The main findings together with the implications for practitioners are discussed in depth.

**Keywords:** Personalized intervention, Deep neural network, Collaborative learning, Knowledge building, Socially shared metacognitive regulation

### 1. Introduction

Computer-supported collaborative learning (CSCL) has been widely adopted in the field of education. CSCL is an effective pedagogical approach that aims to foster the social nature of learning (Jeong, Hmelo-Silver, & Jo, 2019), to co-construct shared understanding and intersubjective meaning making (Stahl, 2006). CSCL is sustained by group interaction to promote socialized learning (Hernández-Sellés, Muñoz-Carril, & González-Sanmamed, 2019). Most research topics in the field of CSCL center on discourse and pattern, factors influencing CSCL, methodology, scripting, scaffolding, and the development of CSCL environments (Tang, Tsai, & Lin, 2014). However, there is still a need to provide personalized intervention in CSCL. To achieve this, it is necessary to automatically analyze the large amount of data generated in CSCL. Previous studies adopted traditional machine learning methods to analyze CSCL data. For example, Mu, Stegmann, Mayfield, Rosé, and Fischer (2012) adopted such methods to automatically segment online discussion transcripts in CSCL. However, conventional machine learning methods depend heavily on human-designed features (Hadi, Al-Radaideh, & Alhawari, 2018) and there is a lack of semantic representations (Shan, Xu, Yang, Jia, & Xiang, 2020), which results in poor performance.

With the rapid development of modern artificial intelligence (AI), AI applications have attracted increasing interest in the field of education (Chen, Xie, & Hwang, 2020). One of the missions of AI in education is to provide personalized guidance, support, or intervention, based on learning status or characteristics (Hwang, Xie, Wah, & Gašević, 2020). However, the provision of personalized intervention to improve learning is still underdeveloped (Hsu, Chiou, Tseng, & Hwang, 2016). Furthermore, although previous studies have exploited conventional machine learning, little work has been done to adopt deep learning technologies in the field of education (Chen, Xie, Zou, & Hwang, 2020). Deep neural networks (DNNs), the type of neural networks used in deep learning, are now able to exceed human accuracy in many fields (Sze, Chen, Yang, & Emer, 2017).

To the best of our knowledge, studies on the real-time analysis of online discussion transcripts gathered during CSCL is very rare, and research on personalized intervention using modern AI techniques in the CSCL context

remain lacking. Additionally, it was found that the use of technology may increase cognitive load (Wu, Huang, Su, Chang, & Lu, 2018). Moreover, the particular intervention could have an impact on socially shared regulation in CSCL context (Lin, 2018). It is very important to investigate the effects of personalized intervention on cognitive load and socially shared metacognitive regulation (SSMR), since few studies to date have examined the issues. Given the scarcity of related studies, this paper proposes a personalized intervention approach based on DNNs and examines the effects of this approach on collaborative knowledge building, group performance, socially shared metacognitive regulation, and cognitive load. The following research questions are addressed:

- (1) Can the personalized intervention approach improve collaborative knowledge building, compared with the conventional online collaborative learning approach?
- (2) Can the personalized intervention approach improve group performance, compared with the conventional online collaborative learning approach?
- (3) Can the personalized intervention approach enhance SSMR, compared with the conventional online collaborative learning approach?
- (4) Can the personalized intervention approach increase cognitive load, compared with the conventional online collaborative learning approach?

## 2. Literature review

### 2.1. Computer-supported collaborative learning

CSCL is concerned with how people learn together with the help of computers (Stahl, Koschmann, & Suthers, 2014). During CSCL, learners communicate and collaborate using digital tools to complete collaborative learning tasks together. CSCL has contributed significantly to enabling learners to acquire knowledge and improve skills (Chen, Wang, Kirschner, & Tsai, 2018). Recently, growing interest was paid to SSMR in CSCL context. SSMR is defined as learners' goal-directed, consensual, and complementary regulation of joint cognitive processes in collaborative learning (Iiskala, Vauras, Lehtinen, & Salonen, 2011). SSMR focused on the metacognitive regulatory episodes at the group level and played a very crucial role in CSCL (De Backer, Van Keer, & Valcke, 2020). Furthermore, CSCL emphasizes the co-construction of knowledge and skills by learners through social interaction (Dillenbourg, 1999; Chen et al., 2018). Therefore, social interaction is the crucial element of collaborative learning (Kreijns, Kirschner, & Jochems, 2003). In the CSCL context, large amounts of data are generated through social interaction, and these data need to be analyzed immediately to provide real-time feedback to learners.

Previous studies have adopted various methods to analyze the data generated during CSCL. For example, social network analysis has often been employed to analyze and visualize the relationships and patterns of interaction in CSCL (Dado & Bodemer, 2017). Epistemic network analysis has been adopted to analyze discourse data to model a cognitive network (Shaffer, Collier, & Ruis, 2016). Furthermore, a social epistemic network signature has been proposed to analyze the social and cognitive dimensions of collaborative learning (Gašević, Joksimović, Eagan, & Shaffer, 2019). In addition, content analysis is a commonly adopted technique for the analysis of discussion transcripts generated in CSCL (Strijbos, Martens, Prins, & Jochems, 2006). Content analysis has often been used to analyze knowledge construction (Gunawardena, Lowe, & Anderson, 1997), cognitive presence (Garrison, Anderson, & Archer, 2001), argumentation (Weinberger, Stegmann, Fischer, & Mandl, 2007), self-regulated learning in collaborative learning (Sobocinski, Malmberg, & Järvelä, 2017), and collective creativity (Tan, Caleon, Jonathan, & Koh, 2014). Moreover, lag sequential analysis has also been employed to analyze behavioural transition (Zheng, Li, Zhang, & Sun, 2019) and temporal differences (Lämsä, Hämäläinen, Koskinen, Viiri, & Mannonen, 2020). However, the aforementioned analysis method was conducted manually to perform lag analysis of discussion transcripts during CSCL. Therefore, it was very difficult to use the lag analysis results to provide real-time feedback and intervention. To progress to a deep understanding of the CSCL process, there is an urgent need to conduct real-time analysis to provide personalized intervention for learners.

### 2.2. Personalized intervention

Learning intervention is conceptualized as the design of supporting strategies and guiding activities to improve learning performance (Zhang, Fei, Quddus, & Davis, 2014). Early learning intervention was employed in the field of special education to provide remedial education for students with learning difficulties (Mesmer & Mesmer, 2008). Subsequently, researchers examined the effects of learning intervention in different learning

settings. For example, Westenskow, Moyer-Packenham, and Child (2017) implemented one-on-one tutoring intervention in the classroom for pupils with low mathematics achievement and found that the intervention produced positive results. Hwang, Chang, Chen, and Chen (2018) engaged students in a four-week mobile learning intervention and found that they outperformed comparable students, in terms of learning achievements and learning motivation. Furthermore, Liu, McKelroy, Corliss, and Carrigan (2017) used the adaptive learning system to implement intervention, and found that adaptive learning intervention contributed to addressing the knowledge gap in chemistry. Hwang, Sung, Chang, and Huang (2020) developed a fuzzy expert system-based to implement adaptive learning intervention through analyzing the learners' cognitive and affective status.

Personalized intervention means that different learners receive different types of intervention, based on their learning status (Zhang, Zou, Miao, Zhang, Hwang, & Zhu, 2020). Early personalized intervention was implemented through instructors' observations. In recent years, the development of learning analytics has increased the power of personalized intervention. Teachers or staff can provide personalized intervention based on the results of learning analytics. For example, Yi et al. (2017) implemented personalized intervention through bulletin messages and email in an online learning environment. Zhang et al. (2020) enacted personalized intervention through individual interviews or sending learning reports, to improve academic performance and learning behaviours in a blended learning setting. Furthermore, Yang, Ogata, Matsui, and Chen (2021) believed that artificial intelligence is shifting from technology to humanity, which means that AI should shift from improving productivity to considering human conditions and having a human-oriented approach. Therefore, personalized intervention should shift from technology-oriented intervention to human-oriented intervention. Previous studies implemented interventions to facilitate collaborative learning through scaffolding (Shin, Kim, & Song, 2020), a digital educational intervention (Männistö et al., 2019) or a metacognitive intervention (Smith & Mancy, 2018). However, very few studies have conducted personalized intervention in the CSCL context. Moreover, there is still a lack of studies on personalized intervention based on modern AI technologies.

### 2.3. Modern artificial intelligence in education

AI can be defined as “computers that mimic cognitive functions that humans associate with the human mind, such as learning and problem-solving” (Russell & Norvig, 2009, p. 2). Traditional AI has usually adopted rule-based or statistical models for prediction (Chen et al., 2020). However, modern AI employs DNN techniques (Yosinski, Clune, Bengio, & Lipson, 2014). Since the development of modern AI, DNNs have been used in many domains, such as natural language processing, speech recognition, image recognition, decision making, and robotics (Hwang et al., 2020).

Typical DNN models include convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory network (LSTM) networks, and bidirectional long short-term memory (BiLSTM) networks. The CNN was proposed by LeCun, Bottou, Bengio, and Haffner (1998) and consists of an input layer, convolution layer, pooling layer, fully connected layer, and output layer. RNNs are designed to deal with (time) sequential data to represent relationships among data points (Schuster & Paliwal, 1997). Based on RNNs, LSTM networks are designed to overcome back-propagation problems; they include an input gate, forget gate, and output gate (Hochreiter & Schmidhuber, 1997). Because of their superior ability to preserve sequence information over time, LSTM networks have obtained strong results in a variety of sequence modelling tasks (Tai, Socher, & Manning, 2015). Furthermore, BiLSTM networks were proposed to overcome the shortcomings of LSTM; a BiLSTM network consists of LSTM units that operate in both directions to analyze the features of the future and the past (Graves & Schmidhuber, 2005).

These DNN models provide the potential for facilitating and optimizing learning in the field of education. For example, Xing and Du (2019) adopted a deep learning algorithm to predict MOOC dropout and provide personalized intervention for at-risk students. Wei, Lin, Yang, and Yu (2017) developed a convolution-LSTM-based model to conduct sentiment analysis of cross-domain MOOC forum postings. Jin, Li, Wang, Zhang, Lin, and Yin (2019) developed a drawing learning system, based on the generative adversarial network, to aid pencil drawing; they found that the system promoted the learners' interest in pencil drawing. Park, Mott, Min, Boyer, Wiebe, and Lester (2019) proposed a multistep deep convolutional generative adversarial network to generate educational game level for computer education. Nevertheless, to the best of our knowledge, very few studies have adopted DNNs in the field of CSCL. It should be noted that DNNs is designed for learning tasks with sequential data and DNNs achieved better performance than traditional machine learning (Prusa & Khoshgoftaar, 2017). Therefore, DNNs is very appropriate for online discussion text classification since the discussion transcripts can be represented as sequences of words. Thus, this study adopted DNNs to provide real-time analysis of online discussion transcripts and personalized intervention in online collaborative learning.

### 3. Personalized intervention based on a deep neural network model

This study evaluated a personalized intervention approach to improve collaborative knowledge building, group performance, and SSMR. This approach included three phases, namely data collection, data analysis, and personalized intervention. Figure 1 shows the framework of the proposed personalized intervention approach. In the first phase, participants completed the online collaborative learning task about computer networks. Figure 2 shows a screenshot of the online collaborative learning platform. All of the participants participated in online collaborative learning through the same platform, which also recorded the online discussion transcripts of all groups. To be noted that only learners in the experimental group can click the button of the latest progress to browse the analysis results.

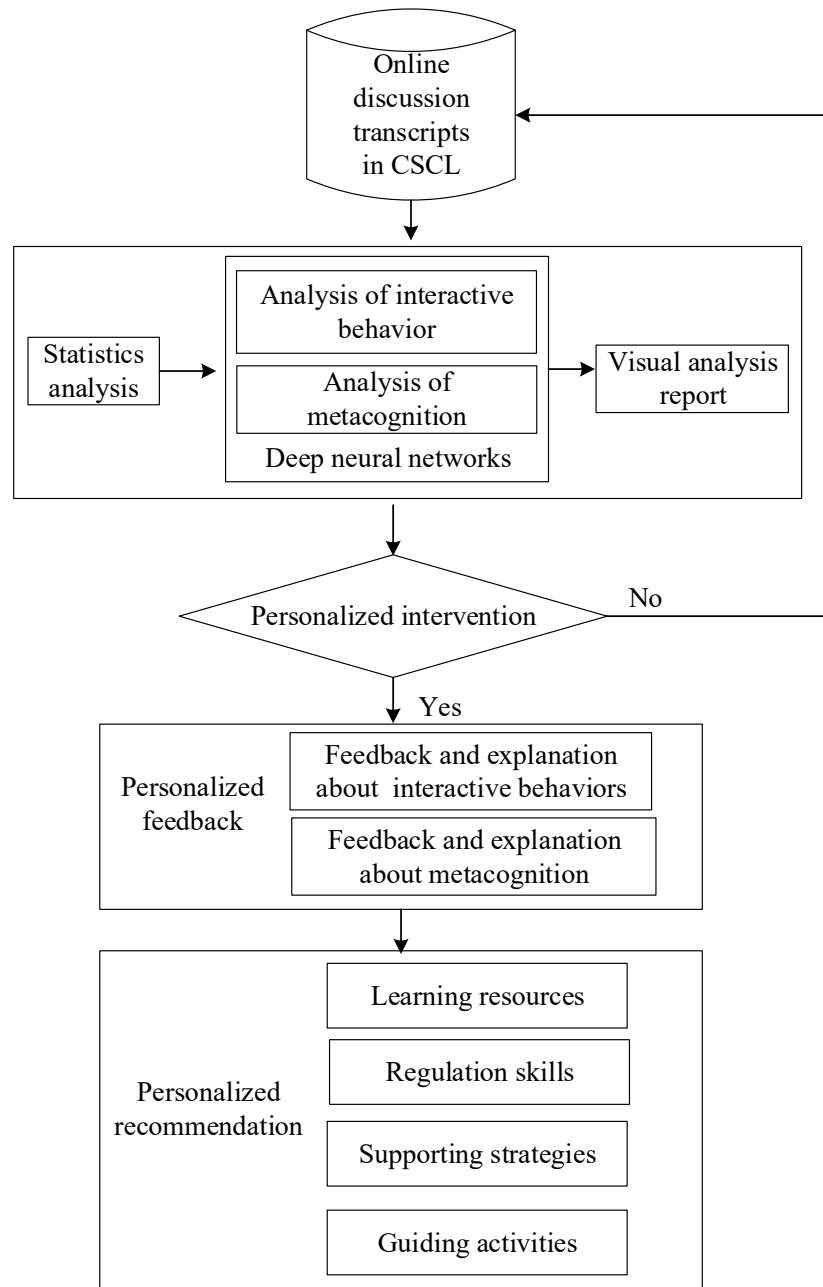


Figure 1. The personalized intervention framework

In the second phase, online discussion transcripts were analyzed in real time through statistical analysis and DNN analysis. The statistical analysis of social interaction included the analysis of the number of posts, duration, interaction frequency, and word cloud. In addition, online interactive behaviors and metacognition of the experimental groups were automatically classified by a DNN model. With regard to interactive behaviors, the online discussion transcripts of the experimental groups were automatically classified into five categories

proposed by authors, namely knowledge building, regulation, support and agreement, asking questions, and off-topic information. With regard to metacognition, the online discussion transcripts were automatically classified into four categories adapted from Zheng (2017), namely planning, monitoring, reflection and evaluation, and off-topic information. The automatic classification results were displayed through a visualization chart and learners could browse at any time. The DNN model was Bidirectional Encoder Representations from Transformers (BERT), which was proposed by Devlin, Chang, Lee, and Toutanova (2019). BERT includes pretraining of deep bidirectional representations and fine-tuning with one additional output layer (Devlin et al., 2019). BERT is trained through the masked language modeling task and independently recovers the masked tokens (Miniae, Kalchbrenner, Cambria, Nikzad, Chenaghlu, & Gao, 2020). In previous studies, BERT achieved the best performance in text classification (González-Carvajal & Garrido-Merchán, 2020). In this study, BERT-Base in Chinese was selected as the pretrained model, 70% of the data were selected as the training set, and 30% were selected as the test set. The parameters were set as follows: the maximum sequence length was 128, the train batch size was 32, the learning rate was 5e-5, and the numbers of train epochs was 3. In addition, other models were used to compare the classification accuracy, as shown in Table 1. It was found that BERT achieved the highest accuracy in terms of interactive behaviors and metacognition classification. Figure 3 shows a screenshot of the statistical result on social interaction and the automatic classification results.



Figure 2. The screenshot of CSCL platform

In the third phase, personalized intervention was provided, based on the analysis results. When the analysis results exceeded the intervention thresholds, our system provided personalized group feedback and recommendations. Personalized group feedback included interactive behaviors and metacognition classification results of each group as well as explanations. For example, when the classification result about interactive behaviors showed that there was off-topic information, the system provided the personalized group feedback “Please focus on the collaborative learning task and don’t discuss off-topic information.” When there was more information about asking questions, the system provided the feedback “Please communicate with your peers to solve problems together. Go ahead!” In addition, when the classification result about metacognition revealed that there was little information about reflection and evaluation, the system provided the feedback “Please reflect and evaluate the collaborative learning progress and group product. Your group can refine the group product further.” Figure 4 shows a screenshot of the personalized group feedback. Moreover, the personalized intervention also provided personalized recommendations and suggestions for learning resources, supporting strategies, and guiding activities. For example, when there were few messages about knowledge building, the system recommended and demonstrated learning materials and knowledge graphs about computer networks. When the classification result about metacognition revealed that there were few messages about planning, the system recommended the construction of a detailed plan about role assignment and scheduling. When the classification result about metacognition revealed that there was little information about monitoring, the system suggested that the group members should monitor and control the collaborative learning process further. Figure 5 and Figure 6 show screenshots of personalized recommendations.

Table 1. The accuracy of different models

Models	Classifications	Accuracy
BERT	Interactive behaviors	0.87
	Metacognition	0.89
LSTM	Interactive behaviors	0.63
	Metacognition	0.85
BiLSTM	Interactive behaviors	0.61
	Metacognition	0.85
Support Vector Machine	Interactive behaviors	0.65
	Metacognition	0.71
Logistic Regression	Interactive behaviors	0.64
	Metacognition	0.76

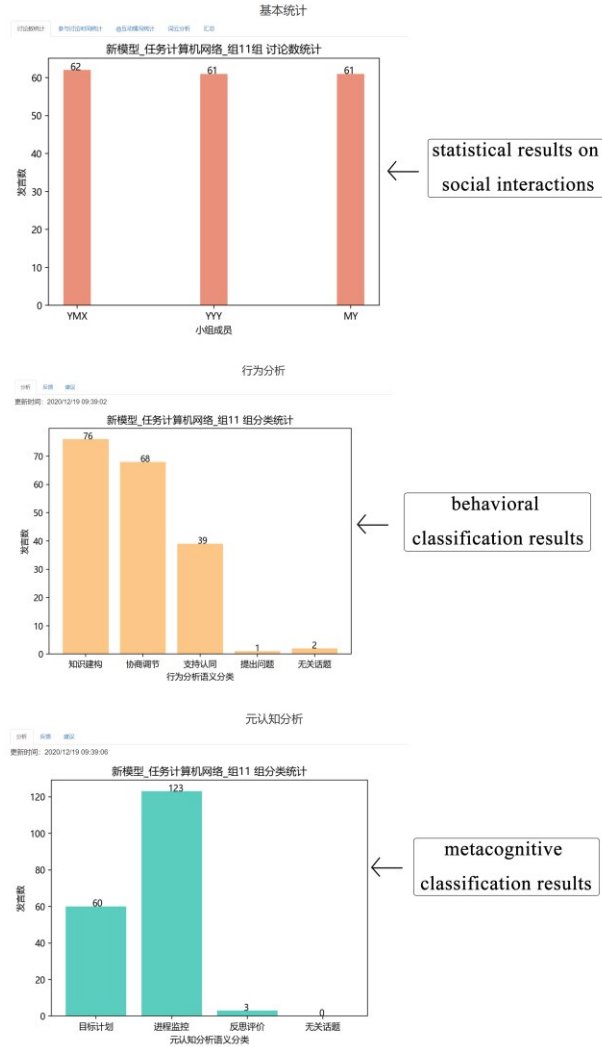


Figure 3. The screenshot of classification results



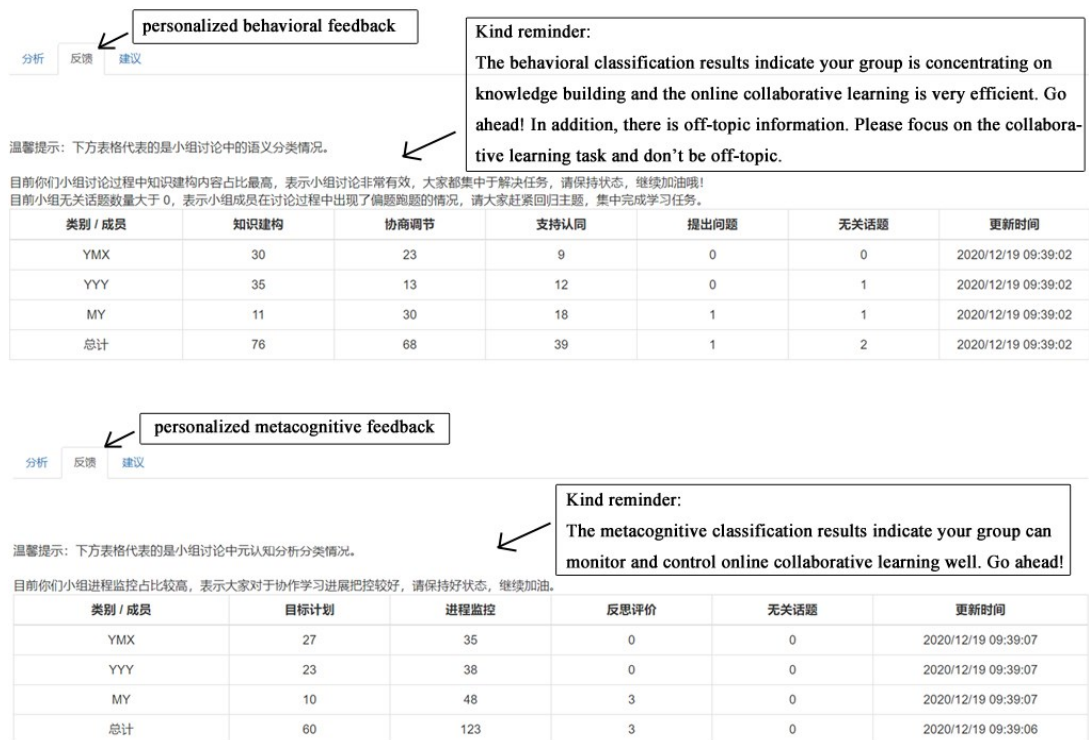


Figure 4. The screenshot of personalized group feedback

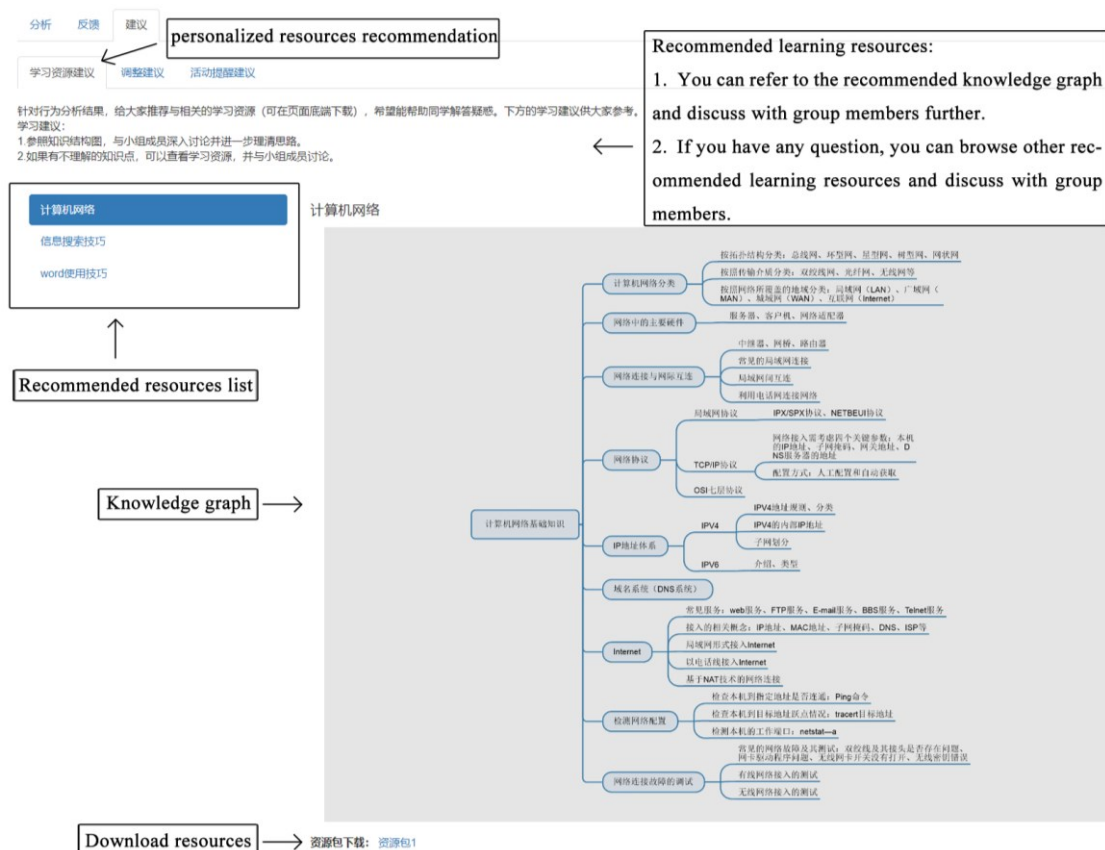


Figure 5. The screenshot of personalized resources recommendations

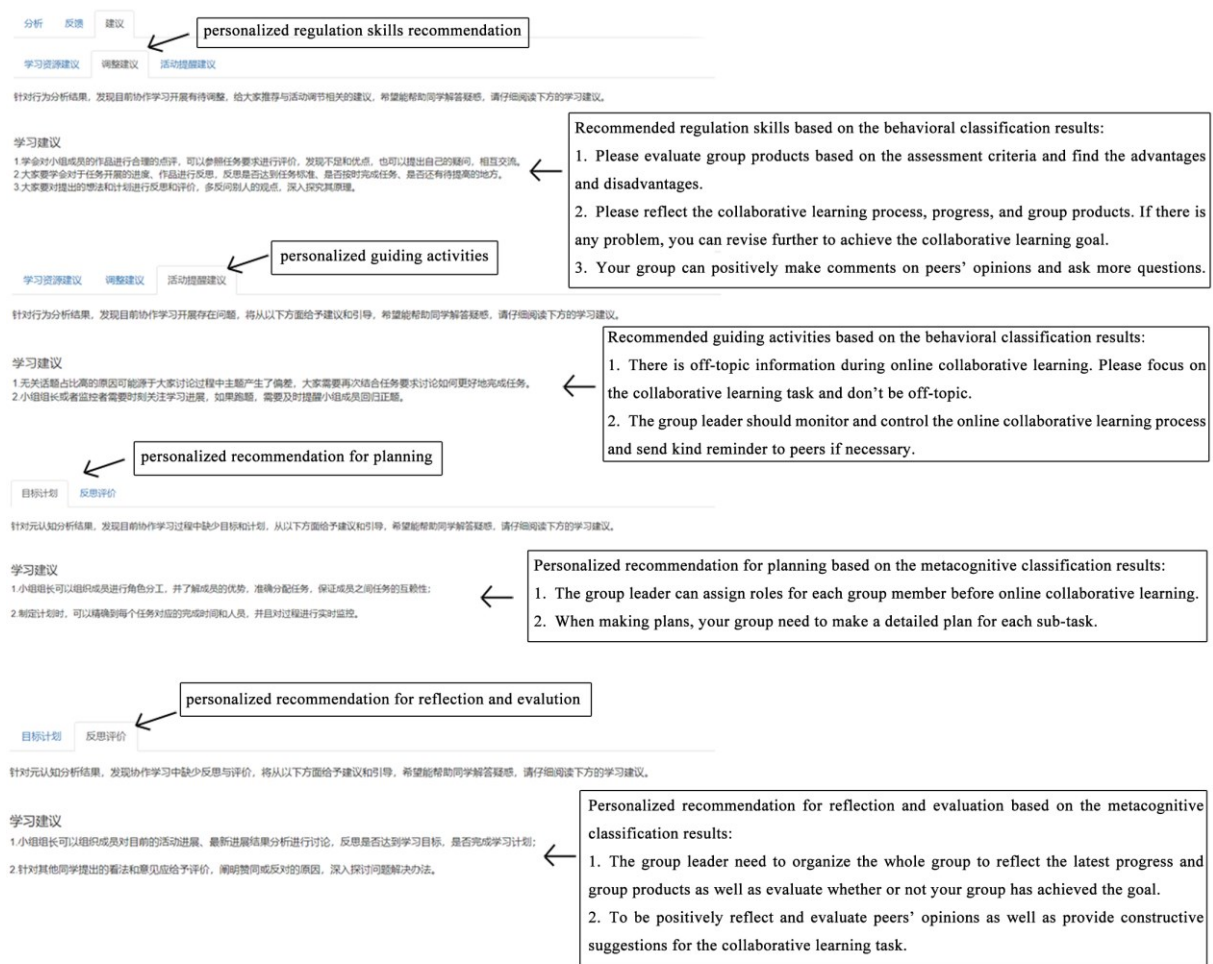


Figure 6. The screenshot of personalized recommendations

## 4. Method

### 4.1. Participants

The participants in the study were from a university in Beijing and were enrolled through posters on campus. Sixty-six college students participated, including eight males and 58 females, with an average age of 21. They majored in education, psychology, history, politics, AI, mathematics, physics, and chemistry, but all participants had prior knowledge about computer networks. All participants were randomly assigned to 11 experimental groups and 11 control groups. Each team contained three students who had not previously collaborated.

### 4.2. Experimental procedure

The experimental procedure included four phases. In the first phase, a pre-test about prior knowledge was conducted. The results of the pre-test indicated that there was no significant difference in prior knowledge between the experimental group and control group ( $t = .68, p = .49$ ). In the second phase, the online collaborative learning platform was introduced and online collaborative learning was conducted for three hours. The experimental group and control group completed the same task, with the same duration. The only difference was that the participants in the experimental group conducted online collaborative learning using the personalized intervention approach, whereas those in the control group used the conventional online collaborative learning approach without personalized intervention. After completing the collaborative learning task, all groups submitted their main ideas and solutions in a Word document, as the group product. In the third phase, a post-questionnaire about cognitive load was completed for 10 minutes. Finally, a semi-structured focus group interview was conducted by two research assistants to understand participants' perceptions of the personalized intervention approach. Six experimental groups were randomly selected and each group participated in a 30-

minute interview in a lab. The interview outline included 10 questions about the personalized intervention approach. The sample interview question included “Do you think the personalized feedback and recommendations were helpful? Why?” The online collaborative learning task was as follows.

With the rapid development of the Internet, college students were encouraged to do pioneering work to serve society. XiaoWang wants to establish a company for online programming education. The first step was to construct a network for the company. Please help XiaoWang to complete the following tasks:

- How should the local network and wireless network be constructed, for the company and for each room? How should the connectivity of the network be tested?
- One day, the local network and wireless network become disconnected. How can they be fixed?
- To overcome fierce market competition, XiaoWang have to investigate the market and competitors. Please help XiaoWang to find and process information about online programming education, by writing a market research report.

#### **4.3. Instruments**

The research instruments included a pre-test and questionnaire about cognitive load. The pre-test aimed to examine whether the experimental group and control group had equivalent prior knowledge about computer networks. The pre-test consisted of 10 single-choice questions, four true-false questions, and three short answer questions with a total score of 100. The example items of the pre-test are “What is computer architecture?” and “Can you list the three applications of computer network?”. The pre-test was developed by the teacher with more than 10 years’ experience of teaching computer course. The pre-test was evaluated by the experienced teacher and a research assistant. The inter-rater reliability using kappa statistics was 0.83, indicating high consistency. This study did not adopt a post-test because collaborative learning performance was measured through the level of collaborative knowledge building and the group products. The cognitive load questionnaire was adapted from Hwang, Yang, and Wang (2013) and it included eight items with a Likert scale: three items that measured mental effort and five items that measured mental load. The reliability of the questionnaire was 0.91 (0.86 for mental load and 0.81 for mental effort). Example items of the cognitive load questionnaire are “The learning content in this learning activity was difficult for me” and “I need to put lots of effort into completing the learning tasks or achieving the learning objectives in this learning activity.”

#### **4.4. Data analysis method**

The data analysis methods include the IIS-map analysis method, content analysis method, and sequential analysis method. To analyze the level of collaborative knowledge building, this study adopted the IIS-map analysis method proposed by Zheng, Yang, and Huang (2012). This method includes three steps, namely drawing the target knowledge graph, coding the online discussion transcripts, and calculating the level of collaborative knowledge building. The collaborative knowledge building level was equal to the activation quantities of all nodes. Two researchers coded the discussion transcripts of 22 groups. The inter-rater reliability using kappa statistics was 0.86, indicating high consistency. SSMR was analyzed based on the coding scheme adapted from Zheng, Li, and Huang (2017), and the analysis unit was a single SSMR episode. Table 2 shows the coding scheme for SSMR. The inter-rater reliability using kappa statistics achieved 0.83, indicating high consistency. The lag sequence analysis method was adopted to analyze the SSMR behavioural transition. The GSEQ 5.1 software developed by Quera, Bakeman, and Gnisci (2007) was employed to conduct behavioural sequence analysis. Moreover, group performance was evaluated, based on the scores of the group products. The assessment criteria were developed by the authors and are shown in Table 3. The inter-rater reliability using kappa statistics was 0.80, indicating high consistency. Finally, face-to-face interviews were recorded by audio and the accuracy of all of the interview data was verified by participants. Content analysis method was used by two research assistants to independently analyze the interview transcripts and group data into inductively categories. Then two assistants reviewed the content and discussed it to come to a consensus when they had conflicts.

Table 2. Coding scheme for socially shared metacognitive regulation

First-level category	Second-level category	Examples
Orienting goals (OG)	Task understanding (TS)	“The tasks require us to find solutions to setting up the local network and wireless network.”
	Setting goals (SG)	“Our group need to complete the three subtasks together.”
Making plans (MP)	Making plans about how to reach the goals, including selecting strategies and setting timelines (MP)	“We need to make a detailed plan about schedule, strategies, and role assignment.”
	Negotiating the division of labor (ND)	“How can we assign roles?”
Enacting strategies (ES)	Advancing and explaining solutions (AE)	“Let’s discuss how to test the connectivity of the network.”
	Coordinating conflicts (CO)	“We have reached a face-saving compromise.”
Monitoring and controlling (MC)	Monitoring or controlling the whole group’s progress (MC)	“How is our group progressing?”
	Claiming (partial) understanding or comprehension failure (CC)	“We have not discovered how to fix the local network.”
	Detecting errors or checking plausibility (DC)	“Our solution is not feasible at all.”
Evaluating and reflecting (ER)	Evaluating current solutions (EV)	“The current solutions still need to be refined further.”
	Reflecting on the group’s goals and progress (RE)	“Our group product is perfect and we have completed the task.”
Adapting metacognition (AP)	Making adaptations to goals, plans, or strategies (MA)	“We have to change our strategies.”

Table 3. Assessment criteria for group product

Dimensions/Rating	16–20	15–11	6–10	1–5
Correctness (20)	Correct opinions and examples.	Correct opinions, but inappropriate examples.	Improper opinions or examples.	Wrong opinions and wrong examples.
Diversity (20)	The solutions and explanations were comprehensive and diverse.	The solutions and explanations were partly comprehensive and diverse.	The solutions and explanations were not diverse.	Solutions and explanations were lacking.
Originality (20)	The solutions and explanations were original and innovative.	The solutions and explanations were partly original.	The solutions and explanations lacked originality.	The solutions and explanations were copied from the Internet.
Completeness (20)	The solutions and explanations were complete and coherent.	The solutions and explanations were complete but not coherent.	The solutions were almost complete, but the explanations were incomplete and incoherent.	Both the solutions and explanations were incomplete.
Format (20)	The Word document was formatted perfectly regarding layout, style, background, color, fonts, type size, and row spacing.	The Word document was formatted well in terms of layout, color, fonts, type size, and row spacing.	The Word document was formatted well only in terms of fonts and type size.	The Word document format was completely disordered.

## 5. Results

### 5.1. Analysis of collaborative knowledge building

This study adopted a one-way ANCOVA (analysis of covariance) to examine whether there were significant differences in collaborative knowledge building between the experimental and control groups. First, the findings of a Kolmogorov–Smirnov test revealed that all datasets were normally distributed ( $p > .05$ ). Second, the assumption of homogeneity of regression was not violated ( $F = 0.01, p = .92$ ). Therefore, the one-way ANCOVA could be performed, with the pre-test as the covariant variable to exclude the effects of pre-test on collaborative knowledge building, the learning approach as the independent variable, and collaborative knowledge building as the dependent variable. Table 4 shows the ANCOVA analysis results. The results revealed a significant difference in collaborative knowledge building between the experimental and control groups ( $F = 12.70, p = .002$ ). Moreover, the mean score of the experimental group was higher than that of the control group. Therefore, the learners who learned with the personalized intervention approach had a higher level of collaborative knowledge building than those who learned with the conventional approach. The eta squared value  $\eta^2 = .40$  indicated a large effect size ( $\eta^2 > .138$ ), according to Cohen (1988). Therefore, the personalized intervention approach had a beneficial effect in increasing the level of collaborative knowledge building. Figure 7 and Figure 8 show the knowledge graphs of an experimental group and control group, respectively. The number besides the node denoted the activation quantity. It is very obvious that the experimental group co-constructed a graph containing more knowledge and relationships.

Table 4. Summary of ANCOVA on collaborative knowledge building

Group	N	Mean	SD	Adjusted mean	SE	F	$\eta^2$
Experimental group	33	385.90	83.56	389.81	31.71	12.70**	.40
Control group	33	232.51	121.27	228.60	31.71		

Note. \*\* $p < .01$ .

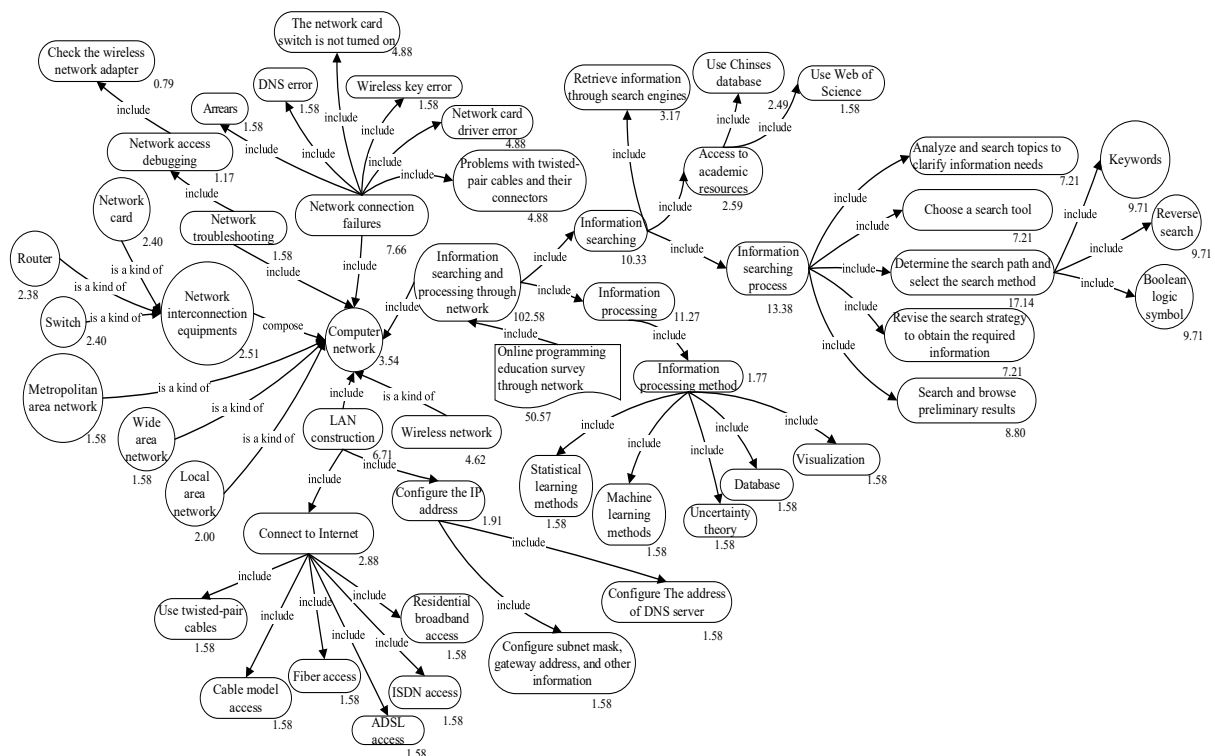


Figure 7. The knowledge graph of an experimental group

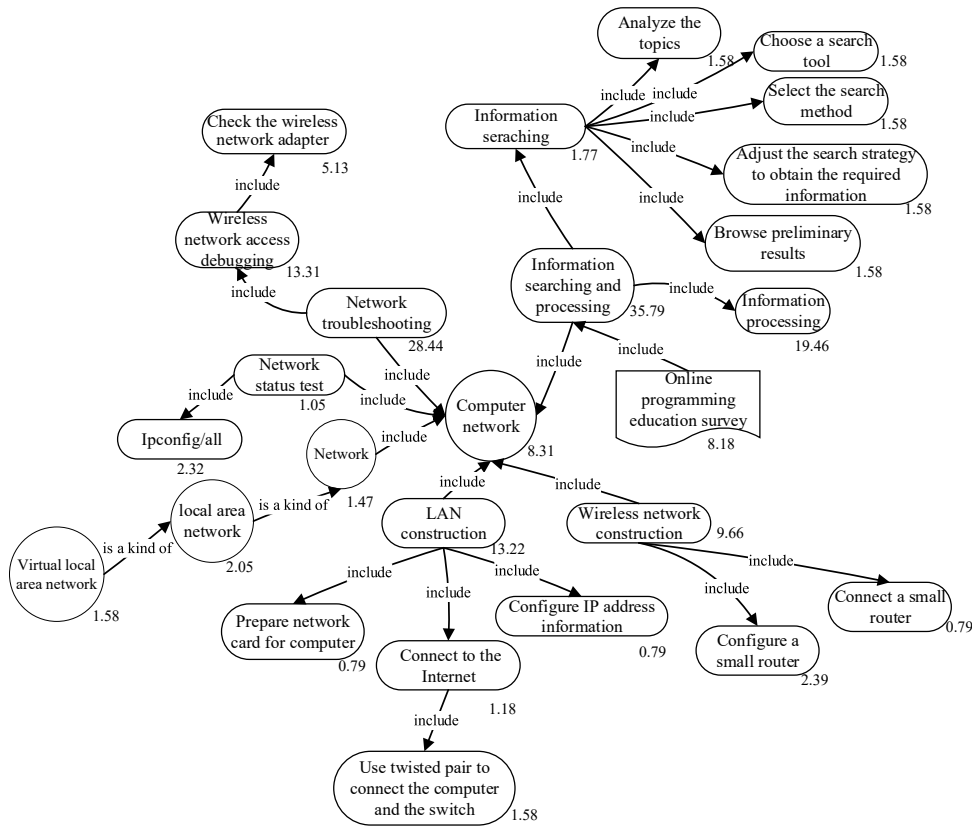


Figure 8. The knowledge graph of a control group

## 5.2. Analysis of group performance

The study also investigated the impacts of the personalized approach on group performance. The scores for the group products were used to evaluate group performance. The results of a Kolmogorov–Smirnov test confirmed that all datasets were normally distributed ( $p > .05$ ). The assumption of homogeneity of regression was not violated ( $F = 1.309$ ,  $p = .268$ ), meaning that the one-way ANCOVA could be performed. As shown in Table 5, there was a significant difference in group performance between the experimental and control groups ( $F = 62.24$ ,  $p = .000$ ). The eta squared value  $\eta^2 = .766$  indicates a large effect size ( $\eta^2 > .138$ ). Therefore, the learners who learned with the personalized intervention approach achieved a higher group performance than those who learned with the conventional approach.

Table 5. Summary of ANCOVA on group performance

Group	N	Mean	SD	Adjusted mean	SE	F	$\eta^2$
Experimental group	33	85.00	4.09	84.90	2.06	62.24***	.766
Control group	33	61.64	8.42	61.73	2.06		

Note. \*\*\* $p < .001$ .

## 5.3. Analysis of socially shared metacognitive regulation

Table 6 shows the descriptive statistics results of SSMR behaviors of the experimental and control groups. The lag sequential analysis method was adopted to analyze the SSMR behavioral transitions. Table 7 shows the results for the experimental group. The vertical direction in Table 7 indicates the starting behaviors and the horizontal direction indicates the subsequent behaviors. The z-score is used to evaluate the possible behavioral sequence transitions. A z-score greater than 1.96 indicates that the behavioral sequence has a significant level (Bakeman & Quera, 2011). As shown in Table 7, there were six significant behavior sequences: OG→MP, MP→MP, MP→ES, ES→MC, ES→AP, and MC→ER. Figure 9 shows the SSMR behavioral transition diagram for the experimental group. In contrast, for the control group, there were only three behavior sequences with a significant level, namely MP→MP, ES→ES, and MC→ER. Table 8 shows the results of the control groups and Figure 10 shows the SSMR behavioral transition diagrams of the control groups. Therefore, the behavioral transitions of the experimental groups were more diverse than those of the control groups. As shown in Table 9,

there were three significant behavior sequences that only occurred in the experimental groups, namely MP→ES, ES→MC, and ES→AP. Therefore, enacting strategies, monitoring and controlling, and adapting metacognition were the crucial regulatory metacognitive behaviors for successful collaborative learning.

Table 6. The descriptive statistics results of SSMR behaviors

		OG	MP	ES	MC	ER	AP
Experimental group	<i>N</i>	10	40	136	127	30	10
	Mean	0.91	3.64	12.36	11.55	2.73	0.91
	<i>SD</i>	1.14	2.11	7.58	7.21	1.95	1.04
Control group	<i>N</i>	5	41	77	110	14	0
	Mean	0.45	3.73	7	10	1.27	0
	<i>SD</i>	0.82	2.83	5.46	6.96	1.19	0

Table 7. Adjusted residuals of the experimental group

Starting behavior	Subsequent behavior					
	OG	MP	ES	MC	ER	AP
Orientating goals (OG)	1.63	2.21*	-1.95	0.23	0.14	-0.56
Making plans (MP)	-1.04	2.36*	2.09*	-1.96	-1.49	-1.17
Enacting strategies (ES)	-0.08	0.06	-2.92	2.39*	-0.26	2.05*
Monitoring and controlling (MC)	0.84	-1.48	1.63	-1.63	2.08*	-1.74
Evaluating and reflecting (ER)	-0.82	-1.04	0.28	1.06	-0.92	0.29
Adapting metacognition (AP)	-0.50	-1.05	1.33	-0.44	-1.00	1.35

Note. \* $p < .05$ .

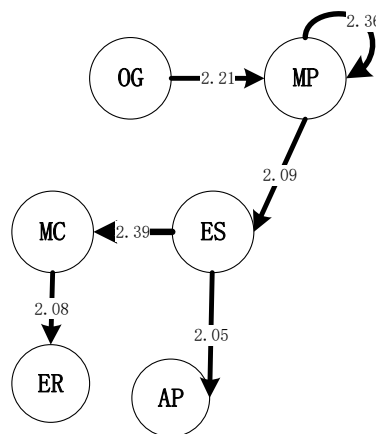


Figure 9. SSMR behavioural transition diagram of the experimental group

Table 8. Adjusted residuals of the control group

Starting behavior	Subsequent behavior					
	OG	MP	ES	MC	ER	AP
Orientating goals (OG)	-0.30	0.36	0.36	-0.24	-0.57	0.00
Making plans (MP)	0.41	2.98*	-1.24	-0.20	-1.77	0.00
Enacting strategies (ES)	-1.40	0.36	2.04*	-1.37	-0.92	0.00
Monitoring and controlling (MC)	1.22	-2.34	-0.72	1.04	2.06*	0.00
Evaluating and reflecting (ER)	-0.35	-1.10	-1.05	1.41	0.95	0.00
Adapting metacognition (AP)	0.00	0.00	0.00	0.00	0.00	0.00

Note. \* $p < .05$ .

Table 9. Significant behaviour sequences that only occurred in the experimental group

Starting behavior	Subsequent behavior					
	OG	MP	ES	MC	ER	AP
Orientating goals (OG)						
Making plans (MP)			MP→ES			
Enacting strategies (ES)				ES→MC	ES→AP	
Monitoring and controlling (MC)						
Evaluating and reflecting (ER)						
Adapting metacognition (AP)						

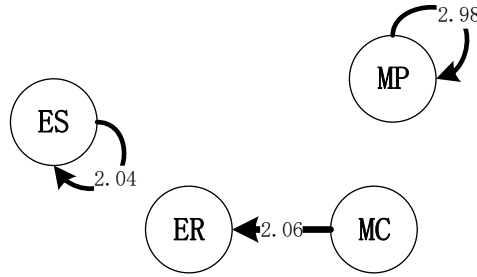


Figure 10. SSMR behavioural transition diagram of the control group

#### 5.4. Cognitive load

The independent-samples  $t$  test was used to examine the difference in cognitive load. As shown in Table 10, there was no significant difference in cognitive load between the experimental and control groups ( $t = 1.50$ ,  $p = .13$ ). Furthermore, there were no significant differences in mental load ( $t = 1.22$ ,  $p = .22$ ) and mental effort ( $t = 0.54$ ,  $p = .58$ ) between the experimental and control groups. Therefore, all of the participants had similar perceptions concerning the collaborative learning tasks. The proposed personalized approach did not increase cognitive load on the participants of the experimental group.

Table 10. Independent sample  $t$ -test results of cognitive load

Dimensions	Group	$N$	Mean	$SD$	$t$
Cognitive load	Experimental group	33	4.50	1.33	1.50
	Control group	33	4.04	1.11	
Mental load	Experimental group	33	4.61	1.35	1.22
	Control group	33	4.23	1.13	
Mental efforts	Experimental group	33	3.20	0.57	0.54
	Control group	33	3.27	0.47	

#### 5.5. Interview results

To gain a better understanding of participants' perceptions of the personalized intervention approach, six experimental groups were randomly selected for interview. The interview data were sorted into three categories. First, all of the interviewees believed that the personalized intervention approach was very helpful for increasing the level of collaborative knowledge building and improving group products. The main reason was that the personalized intervention approach could automatically classify discussion transcripts and provide personalized service based on the analysis results. Learners could keep track of the status and progress of collaborative learning by checking the analysis results. For example, one interviewee stated, "When our group check the latest progress and find that there is many off-topic information, we immediately go back to the collaborative learning task and build knowledge together." Another interviewee stated, "The feedback and suggestions are very helpful. The suggestions for learning resources and guiding activities contributed to our co-constructing knowledge together. We really appreciate it."

Second, all of the interviewees believed that the personalized intervention approach contributed to SSMR. The analysis results on interactive behaviors and metacognition informed learners in the experimental groups to regulate themselves. For example, one interviewee told us, "The analysis results on metacognition show that there is little information about reflection and evaluation. The system reminds us to reflect further on the collaborative learning process and the group product." Another interviewee said, "The metacognition classification results are helpful for SSMR. When our group finds the metacognition status of each group member, we can regulate ourselves immediately, based on the results."

Third, all of the interviewees believed that the personalized intervention approach did not increase cognitive load. For example, one interviewee said, "Our group members like to check the classification results to learn more about the collaborative learning progress. We really need it to regulate ourselves. There is no cognitive load." Another interviewee stated, "The personalized group feedback and suggestions are really necessary and we like to check them when we need. There is no cognitive load for us."



## **6. Discussion**

This study examined the effects of the personalized intervention approach on collaborative knowledge building, group performance, SSMR, and cognitive load in CSCL. The personalized intervention approach was implemented automatically, based on the classification results performed by BERT. The results of the quasi-experiment indicated that the proposed personalized intervention approach significantly improved collaborative knowledge building, group performance, and SSMR behaviours. In addition, it did not increase learners' cognitive load.

### **6.1. Effects on collaborative knowledge building and group performance**

The results of the ANCOVA analysis revealed that learners in the experimental groups outperformed those of the control groups in terms of collaborative knowledge building and group performance. This finding indicates that the personalized intervention approach can efficiently increase the level of collaborative knowledge building and improve the group products. There are several possible explanations for the findings. First, the personalized intervention approach performed the automatic classification of online interactive behaviours, which provided extra information about the progress of online collaborative learning. The classification of online interactive behaviours (showing the numbers of knowledge building, regulation, support, asking questions, and off-topic information) stimulated learners to co-construct knowledge in depth. When learners found that there was off-topic information, they would immediately return to collaborative knowledge building and complete the group products. In addition, the statistical results on social interaction also quantified the contribution of each group member, thereby increasing the group awareness of the members' status. As Yilmaz and Yilmaz (2020) concluded, increasing group awareness contributed to improving knowledge building.

Second, the personalized intervention approach provided personalized group feedback and explanations for each group. The formative feedback and explanations about online interactive behaviours and metacognition helped learners to gain a better understanding of the collaborative learning progress and problems. As Resendes, Scardamalia, Bereiter, Chen, and Halewood (2015) suggested, formative feedback promoted discussion moves to advance knowledge building. Furthermore, the support of the personalized intervention approach increased the sense of collective cognitive responsibility to ensure that the collaborative knowledge building and group products improved (Zhang, Scardamalia, Reeve, & Messina, 2009). Third, the personalized intervention approach provided individualized recommendations for each group. These suggestions, which included various types of learning resources, cases, support strategies, and guiding activities, improved the collaborative knowledge building and group products.

### **6.2. Effects on socially shared metacognitive regulation**

This study found that the personalized intervention approach promoted SSMR behaviours. Learners who used the personalized intervention approach demonstrated more SSMR behaviours than those in the control groups. In addition, the study found that enacting strategies, monitoring and controlling, and adapting metacognition were the critical behaviours for promoting SSMR. There are several possible explanations for these findings. First, the metacognition classification results showed the numbers of planning, monitoring, and reflection and evaluation behaviours during collaborative learning, thereby directly promoting SSMR at the group level. Second, the statistical analysis of social interaction and the classification of interactive behaviours also contributed to SSMR. For example, when group members found that there was little interaction, they would increase interaction with peers. Third, personalized group feedback and recommendation further facilitated group metacognitive regulation and behavioural transition. This finding is consistent with that of De Backer, Van Keer, and Valcke (2016), who believed that feedback promoted groups' metacognitive regulation.

### **6.3. Effects on cognitive load**

The study found that the proposed personalized intervention approach did not increase cognitive load for learners in the experimental group. Learners from the experimental group did not report feeling stressed when the personalized intervention was provided to support collaborative learning. The reason may be that learners checked the latest progress and personalized intervention only when they needed. Furthermore, the personalized intervention was considered very helpful for completing collaborative learning tasks. As Paas, Renkl, and Sweller (2003) revealed that learners' cognitive load can be controlled and reduced by using an effective

instructional design. In addition, learners in the two groups completed the same collaborative learning task, with the same duration. Therefore, there was no significant difference in cognitive load between the experimental and control groups.

#### **6.4. Implications**

The rapid development of AI enables real-time analysis and personalized intervention to improve the performance of collaborative learning. The current study adopted a DNN model to automatically classify online collaborative learning transcripts and provide personalized intervention for each group. This study has several implications for teachers, developers, and practitioners.

First, teachers should provide personalized intervention to improve the performance of collaborative learning. With the aid of AI technology, data generated in online collaborative learning can immediately be analyzed automatically to provide personalized intervention. Types of intervention include supporting strategies, guiding activities, and recommended learning sources. Teachers or practitioners can also evaluate the impacts of personalized intervention on learning performance and perceptions. However, it should be noted that personalized intervention needs to be elaborately designed to achieve the desired effects (Liu et al., 2017).

Second, teachers and practitioners should pay attention to SSMR to achieve productive collaborative learning. It has been found that SSMR is positively related to learning performance (De Backer, Van Keer, & Valcke, 2020). Because learners may have difficulties with SSMR, teachers and practitioners can provide necessary training about SSMR skills before collaborative learning. For example, to improve SSMR, training should be provided in monitoring and controlling collaborative learning processes, as well as adapting metacognition.

Third, researchers and developers need to focus on the latest AI techniques to improve the accuracy of DNN models. For example, more work is required on enhancing the performance of BERT. Increasing training datasets also contributes to improving the accuracy of DNN models (Hestness et al., 2017). Fine-tuning strategies can be adopted to obtain optimized models that achieve better performance. In addition, developers should also develop new DNN models to be applied in different domains.

#### **7. Conclusions**

This study examined the effects of personalized intervention on collaborative knowledge building, group performance, SSMR, and cognitive load. The personalized intervention approach included automatic analysis of interactive behaviors and metacognition, providing personalized group feedback, and providing personalized recommendations. The findings revealed that the proposed personalized intervention approach significantly improved collaborative knowledge building, group products, and SSMR. The study highlighted the contributions of DNNs to providing real-time analysis and personalized intervention in CSCL. The main contribution of the study was to adopt a DNN model to implement personalized intervention in CSCL. The study broadened the understanding of how teachers and practitioners can be guided to provide personalized intervention in CSCL.

The study had several limitations and its results should be generalized with caution. First, the sample size was not large. Future studies will increase the sample size and datasets to improve the accuracy of the model and validate the proposed approach to personalized intervention. Second, the duration of the experiment was short. Future studies will conduct long-term experiments to provide powerful evidence about the personalized intervention approach. Third, the study examined the effects of the personalized intervention approach only on collaborative knowledge building, group performance, and SSMR. Future studies will examine the effects on other variables, such as collective efficacy, problem solving skills, and higher-order thinking skills.

#### **Acknowledgement**

This study is funded by the National Natural Science Foundation of China (61907003).

## References

- Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences*. Cambridge, U K: Cambridge University Press.
- Chen, J., Wang, M., Kirschner, P. A., & Tsai, C. C. (2018). The Role of collaboration, computer use, learning environments, and supporting strategies in CSCL: A meta-analysis. *Review of Educational Research*, 88(6), 799–843. doi:10.3102/0034654318791584
- Chen, X., Xie, H., & Hwang, G. J. (2020). A Multi-perspective study on artificial intelligence in education: Grants, conferences, journals, software tools, institutions, and researchers. *Computers and Education: Artificial Intelligence*, 1, 100005. doi:10.1016/j.caeai.2020.100005
- Chen, X., Xie, H., Zou, D. & Hwang, G.-J. (2020). Application and theory gaps during the rise of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 1, 100002. doi:10.1016/j.caeai.2020.100002
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.
- Dado, M. & Bodemer, D. (2017). A Review of methodological applications of social network analysis in computer-supported collaborative learning. *Educational Research Review*, 22, 159–180. doi:10.1016/j.edurev.2017.08.005
- De Backer, L., Van Keer, H. & Valcke, M. (2016). Eliciting reciprocal peer-tutoring groups' metacognitive regulation through structuring and problematizing scaffolds. *The Journal of Experimental Education*, 84 (4), 804–828. doi:10.1080/00220973.2015.1134419
- De Backer, L., Van Keer, H. & Valcke, M. (2020). Variations in socially shared metacognitive regulation and their relation with university students' performance. *Metacognition and Learning*, 15 (2), 233–259. doi:10.1007/s11409-020-09229-5
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171–4186). Minnesota. Minneapolis. Retrieved from <https://arxiv.org/abs/1810.04805>
- Dillenbourg, P. (1999). What do you mean by collaborative learning. In P. Dillenbourg (Ed.), *Collaborative-learning: Cognitive and Computational Approaches* (pp. 1–19). Oxford, England: Elsevier.
- Garrison, D. R., Anderson, T. & Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of Distance Education*, 15 (1), 7–23. doi:10.1080/08923640109527071
- Gašević, D., Joksimović, S., Eagan, B. R. & Shaffer, D. W. (2019). SENS: Network analytics to combine social and cognitive perspectives of collaborative learning. *Computers in Human Behavior*, 92, 562–577. doi:10.1016/j.chb.2018.07.003
- González-Carvajal, S., & Garrido-Merchán, E. C. (2020). Comparing BERT against traditional machine learning text classification. Retrieved from <https://arxiv.org/pdf/2005.13012.pdf>
- Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings of the 2005 IEEE International Joint Conference on Neural Networks* (Vol. 4, pp. 2047–2052). Montréal, Canada. doi:10.1109/ijcnn.2005.1556215
- 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. doi:10.2190/7mqv-x9uj-c7q3-nrag
- Hadi, W., Al-Radaideh, Q. A. & Alhawari, S. (2018). Integrating associative rule-based classification with Naïve Bayes for text classification. *Applied Soft Computing*, 69, 344–356. doi:10.1016/j.asoc.2018.04.056
- Hernández-Sellés, N., Pablo-César Muñoz-Carril & González-Sanmamed, M. (2019). Computer-supported collaborative learning: An analysis of the relationship between interaction, emotional support and online collaborative tools. *Computers & Education*, 138, 1–12. doi:10.1016/j.compedu.2019.04.012
- Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Ali Patwary, M. M., Yang, Y., & Zhou, Y. (2017). Deep learning scaling is predictable, empirically. Retrieved from <https://arxiv.org/pdf/1712.00409.pdf>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. doi:10.1162/neco.1997.9.8.1735
- Hsu, T.-Y., Chiou, C.-K., Tseng, J. C. R. & Hwang, G.-J. (2016). Development and evaluation of an active learning support system for context-aware ubiquitous learning. *IEEE transactions on learning technologies*, 9 (1), 37–45. doi:10.1109/tlt.2015.2439683

- Hwang, G.-J., Chang, S.-C., Chen, P.-Y. & Chen, X.-Y. (2018). Effects of integrating an active learning-promoting mechanism into location-based real-world learning environments on students' learning performances and behaviors. *Educational Technology Research and Development*, 66 (2), 451–474. doi:10.1007/s11423-017-9567-5
- Hwang, G. J., Sung, H. Y., Chang, S. C., & Huang, X. C. (2020). A Fuzzy expert system-based adaptive learning approach to improving students' learning performances by considering affective and cognitive factors. *Computers and Education: Artificial Intelligence*, 1, 100003. doi:10.1016/j.caeai.2020.100003
- Hwang, G.-J., Xie, H., Wah, B. W. & Gašević, D. (2020). Vision, challenges, roles and research issues of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 1, 100001. doi:10.1016/j.caeai.2020.100001
- 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. doi:10.1016/j.compedu.2013.07.008
- Iiskala, T., Vauras, M., Lehtinen, E. & Salonen, P. (2011). Socially shared metacognition of dyads of pupils in collaborative mathematical problem-solving processes. *Learning and Instruction*, 21 (3), 379–393. doi:10.1016/j.learninstruc.2010.05.002
- Jeong, H., Hmelo-Silver, C. E. & Jo, K. (2019). Ten years of computer-supported collaborative learning: A Meta-analysis of CSCL in STEM education during 2005–2014. *Educational Research Review*, 28, 100284. doi:10.1016/j.edurev.2019.100284
- Jin, Y., Li, P., Wang, W., Zhang, S., Lin, D., & Yin, C. (2019). GAN-based pencil drawing learning system for art education on large-scale image datasets with learning analytics. *Interactive Learning Environments*, 1–18. doi:10.1080/10494820.2019.1636827
- Kreijns, K., Kirschner, P. A. & Jochems, W. (2003). Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: A Review of the research. *Computers in Human Behavior*, 19(3), 335–353. doi:10.1016/s0747-5632(02)00057-2
- Lämsä, J., Hämäläinen, R., Koskinen, P., Viiri, J. & Mannonen, J. (2020). The Potential of temporal analysis: Combining log data and lag sequential analysis to investigate temporal differences between scaffolded and non-scaffolded group inquiry-based learning processes. *Computers & Education*, 143, 103674. doi:10.1016/j.compedu.2019.103674
- Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. doi:10.1109/5.726791
- Lin, J. W. (2018). Effects of an online team project-based learning environment with group awareness and peer evaluation on socially shared regulation of learning and self-regulated learning. *Behaviour & Information Technology*, 37(5), 445–461. doi:10.1080/0144929X.2018.1451558
- Liu, M., Mckelroy, E., Corliss, S. B. & Carrigan, J. (2017). Investigating the effect of an adaptive learning intervention on students' learning. *Educational Technology Research and Development*, 65 (6), 1605–1625. doi:10.1007/s11423-017-9542-1
- Männistö, M., Mikkonen, K., Vuopala, E., Kuivila, H. M., Virtanen, M., Kyngäs, H., & Kääriäinen, M. (2019). Effects of a digital educational intervention on collaborative learning in nursing education: A Quasi-experimental study. *Nordic Journal of Nursing Research*, 39(4), 191–200. doi:10.1177/2057158519861041
- Mesmer, E. M. & Mesmer, H. A. E. (2008). Response to Intervention (RTI): What teachers of reading need to know. *The Reading Teacher*, 62 (4), 280–290. doi:10.1598/rt.62.4.1
- Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2020). Deep learning based text classification: A Comprehensive review. *ACM Computing Surveys*, 54(3), 1–40. doi:10.1145/3439726
- Mu, J., Stegmann, K., Mayfield, E., Rosé, C. & Fischer, F. (2012). The ACODEA framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions. *International Journal of Computer-Supported Collaborative Learning*, 7(2), 285–305. doi:10.1007/s11412-012-9147-y
- Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational psychologist*, 38(1), 1–4. doi:10.1207/S15326985EP3801\_1
- Park, K., Mott, B. W., Min, W., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2019). Generating educational game levels with multistep deep convolutional generative adversarial networks. In *Proceedings of 2019 IEEE Conference on Games (CoG)* (pp. 1–8). United Kingdom, London. doi: 10.1109/CIG.2019.8848085
- Prusa, J. D., & Khoshgoftaar, T. M. (2017). Improving deep neural network design with new text data representations. *Journal of Big Data*, 4: 7. doi:10.1186/s40537-017-0065-8
- Quera, V., Bakeman, R. & Gnisci, A. (2007). Observer agreement for event sequences: Methods and software for sequence alignment and reliability estimates. *Behavior Research Methods*, 39 (1), 39–49. doi:10.3758/bf03192842
- Resendes, M., Scardamalia, M., Bereiter, C., Chen, B. & Halewood, C. (2015). Group-level formative feedback and metadiscourse. *International Journal of Computer-Supported Collaborative Learning*, 10 (3), 309–336. doi:10.1007/s11412-015-9219-x

- Russell, S. J., & Norvig, P. (2009). *Artificial Intelligence: A Modern approach* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.
- Schuster, M. & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11), 2673–2681. doi:10.1109/78.650093
- Shaffer, D. W., Collier, W., & Ruis, A. R. (2016). A Tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics*, 3(3), 9–45. doi: 10.18608/jla.2016.33.3
- Shan, G., Xu, S., Yang, L., Jia, S. & Xiang, Y. (2020). Learn#: A Novel incremental learning method for text classification. *Expert Systems with Applications*, 147, 113198. doi:10.1016/j.eswa.2020.113198
- Shin, Y., Kim, D., & Song, D. (2020). Types and timing of scaffolding to promote meaningful peer interaction and increase learning performance in computer-supported collaborative learning environments. *Journal of Educational Computing Research*, 58(3), 640–661. doi:10.1177/0735633119877134
- Smith, J. M., & Mancy, R. (2018). Exploring the relationship between metacognitive and collaborative talk during group mathematical problem-solving—what do we mean by collaborative metacognition? *Research in Mathematics Education*, 20(1), 14–36. doi:10.1080/14794802.2017.1410215
- Sobocinski, M., Malmberg, J. & Järvelä, S. (2017). Exploring temporal sequences of regulatory phases and associated interactions in low- and high-challenge collaborative learning sessions. *Metacognition and Learning*, 12(2), 275–294. doi:10.1007/s11409-016-9167-5
- Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge*. Cambridge, MA: MIT Press.
- Stahl, G., Koschmann, T., & Suthers, D. (2014). Computer-supported collaborative learning. In R. K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (pp. 479–500). Cambridge University Press. doi:10.1017/CBO9781139519526.029
- Strijbos, J.-W., Martens, R. L., Prins, F. J. & Jochems, W. M. G. (2006). Content analysis: What are they talking about? *Computers & Education*, 46(1), 29–48. doi:10.1016/j.compedu.2005.04.002
- Sze, V., Chen, Y.-H., Yang, T.-J. & Emer, J. S. (2017). Efficient processing of deep neural networks: A Tutorial and survey. *Proceedings of the IEEE*, 105(12), 2295–2329. doi:10.1109/jproc.2017.2761740
- Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. Retrieved from <https://www.aclweb.org/anthology/P15-1150.pdf>
- Tan, J. P.-L., Caleon, I. S., Jonathan, C. R., & Koh, E. (2014). A Dialogic framework for assessing collective creativity in computer-supported collaborative problem-solving tasks. *Research and Practice in Technology Enhanced Learning*, 9(3), 411–437.
- Tang, K.-Y., Tsai, C.-C. & Lin, T.-C. (2014). Contemporary intellectual structure of CSCL research (2006–2013): A Co-citation network analysis with an education focus. *International Journal of Computer-Supported Collaborative Learning*, 9(3), 335–363. doi:10.1007/s11412-014-9196-5
- Wei, X., Lin, H., Yang, L. & Yu, Y. (2017). A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification. *Information*, 8(3), 92. doi:10.3390/info8030092
- Weinberger, A., Stegmann, K., Fischer, F., & Mandl, H. (2007). Scripting argumentative knowledge construction in computer-supported learning environments. In F. Fischer, I. Kollar, H. Mandl, & J. M. Haake (Eds.), *Scripting Computer-Supported Collaborative Learning* (pp. 191–211). doi:10.1007/978-0-387-36949-5\_12
- Westenskow, A., Moyer-Packenham, P. S., & Child, B. (2017). An Iceberg model for improving mathematical understanding and mindset or disposition: An Individualized summer intervention program. *Journal of Education*, 197(1), 1–9. doi:10.1177/002205741719700102
- Wu, T.-T., Huang, Y.-M., Su, C.-Y., Chang, L., & Lu, Y. C. (2018). Application and analysis of a mobile e-book system based on project-based learning in community health nursing practice courses. *Educational Technology & Society*, 21(4), 143–156.
- Xing, W., & Du, D. (2019). Dropout prediction in MOOCs: Using deep learning for personalized intervention. *Journal of Educational Computing Research*, 57(3), 547–570. doi:10.1177/0735633118757015
- Yang, S. J., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 2, 100008. doi:10.1016/j.caeai.2021.100008.
- Yi, B., Zhang, D., Wang, Y., Liu, H., Zhang, Z., Shu, J., & Lv, Y. (2017). Research on personalized learning model under informatization environment. In *2017 International Symposium on Educational Technology (ISET)* (pp. 48–52). IEEE. Retrieved from <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&number=8005386>

- Yilmaz, R. & Yilmaz, F. G. K. (2020). Examination of the effectiveness of the task and group awareness support system used for computer-supported collaborative learning. *Educational Technology Research and Development*, 68(3), 1355–1380. doi:10.1007/s11423-020-09741-0
- Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Proceedings of the 27th International Conference on Neural Information Processing Systems* (Vol. 2, pp. 3320–3328). Retrieved from <https://proceedings.neurips.cc/paper/2014/file/375c71349b295fbc2dedca9206f20a06-Paper.pdf>
- Zhang, J. H., Zou, L. C., Miao, J. J., Zhang, Y. X., Hwang, G. J., & Zhu, Y. (2020). An Individualized intervention approach to improving university students' learning performance and interactive behaviors in a blended learning environment. *Interactive Learning Environments*, 28(2), 231–245. doi:10.1080/10494820.2019.1636078
- Zhang, J., Scardamalia, M., Reeve, R., & Messina, R. (2009). Designs for collective cognitive responsibility in knowledge-building communities. *The Journal of the learning sciences*, 18(1), 7–44. doi:10.1080/10508400802581676
- Zhang, Y., Fei, Q., Quddus, M., & Davis, C. (2014). An Examination of the impact of early intervention on learning outcomes of at-risk students. *Research in Higher Education Journal*, 26, 1–12.
- Zheng, L. (2017). *Knowledge building and regulation in computer-supported collaborative learning*. Singapore: Springer.
- Zheng, L., Li, X., & Huang, R. (2017). The Effect of socially shared regulation approach on learning performance in computer-supported collaborative learning. *Educational Technology & Society*, 20(4), 35–46.
- Zheng, L., Li, X., Zhang, X., & Sun, W. (2019). The Effects of group metacognitive scaffolding on group metacognitive behaviors, group performance, and cognitive load in computer-supported collaborative learning. *The Internet and Higher Education*, 42, 13–24. doi:10.1016/j.iheduc.2019.03.002
- Zheng, L., Yang, K., & Huang, R. (2012). Analyzing interactions by an IIS-Map-based method in face-to-face collaborative learning: An Empirical study. *Educational Technology & Society*, 15(3), 116–132.

## Teachable Agent Improves Affect Regulation: Evidence from Betty's Brain

Jian-Hua Han<sup>1\*</sup>, Keith Shubeck<sup>2,3</sup>, Geng-Hu Shi<sup>2,3</sup>, Xiang-En Hu<sup>2,3,4,5</sup>, Lei Yang<sup>4</sup>, Li-Jia Wang<sup>3,5</sup>, Wei Zhao<sup>1</sup>, Qiang Jiang<sup>1</sup> and Gautum Biswas<sup>6</sup>

<sup>1</sup>School of Information Science and Technology, Northeast Normal University, China // <sup>2</sup>Department of Psychology, The University of Memphis, USA // <sup>3</sup>Institute for Intelligent Systems, The University of Memphis, USA // <sup>4</sup>School of Psychology, Central China Normal University, China // <sup>5</sup>Electrical and Computer Engineering, The University of Memphis, USA // <sup>6</sup>Institute for Software Integrated Systems, Vanderbilt University, USA // hanjh675@nenu.edu.cn // keithshubeck@gmail.com // genghushi@gmail.com // xiangenhu@gmail.com // yungly@mails.ccnu.edu.cn // wlj54188@gmail.com // zhaow577@nenu.edu.cn // jiangqiang@nenu.edu.cn // gautam.biswas@vanderbilt.edu

\*Corresponding author

**ABSTRACT:** Intelligent learning technologies are often applied within the educational industries. While these technologies can be used to create learning experiences tailored to an individual student, they cannot address students' affect accurately and quickly during the learning process. This paper focuses on two core research questions. How do students regulate affect and what are the processes that affect regulation? First, this paper reviews the affect regulation methods and processes in an intelligent learning environment based on affective transition and affect compensation. This process, along with affect analysis, affect regulation, intelligent agents, and an intervention strategy can be used to analyze specific affect regulation methods and improve the affective regulation system. Seventy-two 7<sup>th</sup> grade students were randomly placed into an experimental condition that used Betty's Brain, an intelligent tutoring system (ITS), or a classroom control. A lag sequence analysis and a multinomial processing tree analysis of video data captured at 25-minute intervals revealed significant differences in affect transitions frequencies between the two groups. Based on the results of the above analyses and after-class interviews, we found that Betty's Brain was able to promote effective affect-regulation strategies to students in the domain of forest ecosystems.

**Keywords:** Teachable agent, Affect, Regulation, Tutoring, Betty's Brain

### 1. Introduction

Learning technologies have been widely used in education, which gradually changed the demand for talent and new educational formats (Liu & Lemeire, 2017). On one hand, these technologies benefit education (Popenici & Kerr, 2017). For example, online learning platforms make it possible for students to learn anytime and anywhere (Du et al., 2019). Recommendation algorithms in ITSs can be used to select adaptive content that fits a student's aptitude, characteristics, and learning progress (Wang et al., 2019). However, learning technologies are not without their disadvantages. For example, students may easily find themselves physically isolated in online learning environments, and they may feel helpless when they encounter difficulties (Raufelder et al., 2018). The status quo of learning technologies is that they make learning content easily accessible but they generally do not improve students' affective well-being. Students often suffer from inattention and lose navigation due to non-adaptive media materials, redundant content, and difficult tasks (Burek, 2017; Lim, 2004). Consequently, many students may disengage from the learning content and have unsatisfactory learning gains. Over time, they may feel fatigue and experience negative affect (Fida et al., 2015; Arsenio & Loria, 2014). Therefore, it is important to investigate the role of affect in technology-based learning environments, like ITSs, and the potential solutions for reducing negative affect and their detriments to learning.

Schutz et al. (2007) pointed out that affect influences students' motivation for learning. Research has shown that students who engage in exciting learning activities experience positive affect and have high learning gains (Gross, 1998). Lu (2012) found that learning activities that make students feel happy are important in teaching. Alkhalaf (2018) found that negative affect might lead to poor academic performance. Academic performance can be improved by increasing positive affect and through continuously combating or managing anxiety during learning. By examining students' degree of concentration, patience and learning willingness, Hwang et al. (2020) found that students using an adaptive learning system with affective and cognitive performance analysis mechanisms had significantly lower levels of mathematical anxiety than those who used the conventional learning system.

An ITS is a computer system that aims to provide immediate and individualized instruction or feedback to learners, usually without intervention from a human teacher (Patrut & Spataru, 2016). ITSs have the potential to

help students manage their affect. For example, web cameras and sensors enable ITSs to capture students' facial expressions and other physiological data that can be converted to affect information (i.e., Kołakowska et al., 2020). Then, an ITS can use various algorithms to provide feedback directly or indirectly to learners about how they can regulate their emotions. To explore the impact of an ITS on students' affect during learning, it is important to determine the mechanism of affect regulation and affect transitions when using an ITS.

## **2. Affect regulation in intelligent tutoring systems**

### **2.1. Affect regulation and recognition**

Affect is a kind of inner reaction of cognitive activity. It greatly impacts an individual's behavior. It can also influence an individual's behavior indirectly through affect reinforcement (Zhang, 2008). The affect regulation process can suppress and weaken negative affect, and can also maintain and enhance positive affect (Gross & John, 2003; Thompspon, 1994). For example, e-learning with affect regulation can significantly improve math performance for students with autism spectrum disorder (Chu et al., 2020). The transition from negative affect to positive affect depends on external feedback and internal regulation. Russell (2003) describes affect as consisting of valence (pleasure to displeasure) and arousal (active to inactive). When plotted, valence increases from left to right along the x-axis, and arousal increases moving upwards on the y-axis (Posner et al., 2005). Generally, affective states relevant to learning include boredom, flow, confusion, frustration, surprise, and delight (Craig et al., 2004). Affect occurs between students' cognitive balance and imbalance between boredom, frustration, confusion, and flow (D'Mello & Graesser, 2012). Boredom has negative valence and low arousal. Flow has positive valence and moderate arousal, whereas confusion has negative valence and moderate arousal. Finally, frustration has negative valence with high arousal (Baker et al., 2010). D'Mello and Graesser (2012) developed a model of affective state transitions based on this concept of equilibrium and disequilibrium by observing the main state transitions that occurred in AutoTutor (Nye, Graesser, & Hu, 2014) sessions. For example, flow may transition to confusion, which may transition to frustration or boredom (D'Mello et al., 2007).

Currently, there are three methods that are typically used to detect affect. For example, affect can be detected by using external devices like cameras, recorders, or other sensors that collect student body expressions (e.g., postures and gestures), facial expressions, verbal expressions (e.g., tone and timbre), and physical and psychological information (e.g., heartbeat, blood pressure and skin conductance). Affect can also be tracked through surveys with various affect scales, including questionnaires, self-reports, observations, and interviews. For example, the User Engagement Survey (UES) is used to measure attention, endurance, and participation (Grafsgaard et al., 2012). The third method is through system analysis (Pentel, 2015). This involves analyzing affect based on student interaction logs, accessing paths, frequency of mouse clicks, duration of staying on the page, and interactions with an ITS. Due to the situational and persistent features of affect, scholars can predict the affect of the next moment using the affective characteristics (e.g., intensity and classification) of the previous moment (Yu et al., 2013). An API can match captured images of an individual with the system model and automatically segment the expression into units, then the program can analyze the affect segmentation points to output affect and features (Maheshwari & Nagendhiran, 2017). By using posture estimation (Grafsgaard et al., 2012) and a gesture detection algorithm (Grafsgaard et al., 2012), a depth image regular pattern can be used to analyze the students' interest and concentration in the learning content. Although the analysis of valence and arousal is an effective method for predicting affect, the affect transition framework (D'Mello et al., 2007; D'Mello & Graesser, 2012) provides essential theoretical support for further exploration affect regulation and its effect on learning.

### **2.2. Affect regulation methods in intelligent tutoring systems**

ITSs have different ways of capturing data relevant to affective states, which can be used to inform future system actions. Some use domain-independent rules (e.g., IF-THEN) and non-independent strategies (e.g., "You have done well"), which are used to achieve affect reinforcement (D'Mello & Graesser, 2013). Some use decision trees and sequential covering algorithms (e.g., AQ, CN2 and PIPPER), which are used to extract dataset rules for learning diagnosis (Quinlan, 1990; D'Mello & Graesser, 2013). Others use probabilistic models, like dynamic decision networks, which can be used to diagnose, evaluate, predict, and determine affect (Conati, 2002). Some are based on affect stratification. For example, the Hidden Markov Model and Baum-Welch algorithm can be used to output state transition probability matrix and vector parameters to evaluate affect (Collins, 1990; Liu & Lemeire, 2017; Thornton & Tamir, 2017). Some use dynamic Bayesian networks to focus on the causes and



effects of affect, and probabilistic frameworks to handle high-level uncertainties to identify affect (Conati, 2002). Some use corpora, latent semantic analysis, word vectors and other analytical texts to predict affect response.

Teaching agents in an ITS can respond to and regulate negative affect by providing appropriate tutoring strategies and feedback. D'Mello et al. (2010) observed postures, facial expressions, and dialogue cues to stimulate pedagogical interventions, regulate boredom, frustration, and confusion, and then promoted participation and task persistence in AutoTutor. Wayang Outpost (Arroyo et al., 2014) adopted heuristic strategies for responding to students' affect, including text information and mapping learning behaviors. Their results showed that students can alleviate their boredom and change their behaviors based on digital interventions (Woolf et al., 2009). Although the students in the experimental and control group showed very similar feelings of pleasure, arousal, and dominance, Daradoumis and Arguedas (2020) found that the experimental group was slightly more expressive about their personal satisfaction through an affect pedagogical agent. Based on the theorized model of D'Mello and Graesser (2012), Alexandra et al. (2019) examined three types of affective transitions and their correlations with pretest-to-posttest learning. They found that the presence of boredom indicates a student's knowledge state, but not their learning. In summary, ITSs are mostly used in one-to-one tutoring simulations of human teachers, and they use domain and student models to support students' cognition and affect regulation.

### **2.3. Affect regulation processes in intelligent tutoring systems**

Qin et al. (2014) built an affect compensation structure, using affect recognition, personalized affect regulation, and negative affect compensation. First, multi-modal methods (e.g., facial expressions, language, behavior, and interactive text) use high resolution cameras and wearable sensors to recognize students' positive affect or negative affect, such as frustration. Next, personalized regulation methods are used to analyze the student's characteristics and regulation strategies, and then judge the affect regulation methods. They then use affect compensation (including expert tutoring and peer help) to enhance the system's confidence in the student's optimal affective states. Affect compensation can optimize the affect database, and the affect database can be used for affect recognition and negative affect compensation. Finally, based on historical compensation cases and compensation lists, the systems can be used to alleviate negative affect. According to the affect compensation structure (Qin et al., 2014) and the affect recognition method, the four functional modules and the affect regulation processes can be implemented in an ITS (see Figure 1).

The first module is an affect analysis module. Affect analysis is key to providing intervention strategies. System tracking can determine whether students have studied or not and can also track their affect and transitions. Affect extraction is defined as using self-report and text mining to extract affective valence and arousal. Affect recognition is based on recording and quantifying personal physiological, psychological, and cognitive information to detect affect states. For example, some scholars use cameras and wearable devices to identify student affect in MetaTutor (Harley et al., 2015). The second module is an intelligent agent module. This mainly involves an intelligent agent, like an expert, a teacher, or a peer, who is a virtual animated character that plays a certain role during an interactive session in an ITS. Expert agents have a wealth of knowledge in various disciplines and domains, such as students communicating with virtual doctors and patients, or reasoning about the patient symptoms of island residents in Crystal Island (Taub et al., 2017). Teacher agents track students' knowledge construction processes, for instance, the agents in AutoTutor that judge questions and then give appropriate feedback by leveraging "expectation-misconception tailored dialogue" (D'Mello & Graesser, 2013). Peer agents in Betty's Brain act as learning peers and assistants by using a learn-by-teaching method to build knowledge (Han et al., 2019). The third module is an affect regulation module. Students can regulate their affect by themselves and can also regulate their affect by external feedback and interventions. The external feedback and interventions are mostly based on an analysis of learning characteristics to achieve precise tracing and interventions with dialogues, student logs, and questionnaires in real-time. Self-report data from students suggest they can acquire appropriate affective and cognitive feedback automatically in Betty's Brain (Biswas et al., 2016). Students can adjust their cognitive affect in real-time based on lists and features selection methods, agent dialogue, problem clues, animation prompts and diagnostic reports (Taub et al., 2017). Sometimes, ITSs can present empathetic agents and virtual companions to augment students' awareness of cognitive presence and affect presence, such as in Wayang Outpost (Arroyo et al., 2014). The fourth module is an intervention strategy module. Intervention strategies are the means and methods of affective intervention. Individual moods and cognitive dilemmas are affected by affect, and some scholars use self-explanations and learning-early-warnings to relieve learner confusion about stress analysis in Andes, an ITS for physics (VanLehn et al., 2010). Peer agents can intervene with students in mathematical problem solving in real time, such as SimStudent agent dialogues (Matsuda et al., 2013). Intelligent systems can provide adaptive resources and suggestions based on cognitive impairments or resources property. Taking Wayang Outpost as an example, the system can provide

cognitive clues and suggestions in addition to different media materials like video, sound, text, and test (Woolf et al., 2009). A system can provide hints and suggestions and help students solve problems correctly in ASSISTments, given the steps and results of students' questions. (Heffernan & Heffernan, 2014).

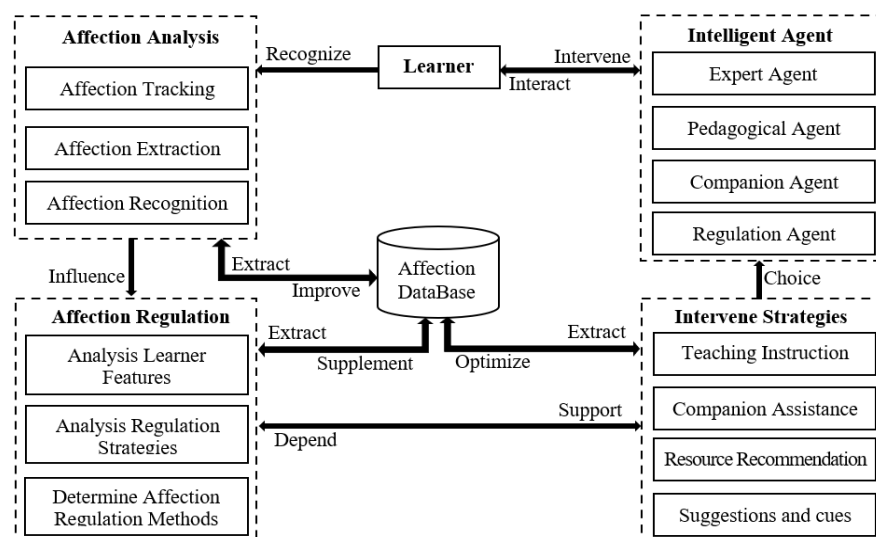


Figure 1. Affective regulation processes in intelligent tutoring environment (adapted from Qin et al., 2014)

In general, based on personal records and known affect, an ITS analyzes resource acceptance manners, preferences, and moods, and uses personalized regulation strategies to increase and decrease positive and negative affect. It can supplement the affect database if the system does not have a student's affect records. Tutoring strategies grounded in ITS cognitive principles and algorithms (i.e., error recognition and correction, student modeling, and natural language dialogue) instruct agents to indirectly address student affect. These strategies are also sent directly to students, which allow computers to act as virtual instructors to impart knowledge and provide adaptive feedback for students. Moreover, based on the frequency of mouse clicks and the path of page access, an ITS analyzes arousal and valence to predict the next affective state, and uses some encouragement, care, praise, and criticism via agents to optimize database adjustment strategies. Additionally, systems can inquire about learner's affect and present them learning tasks and their progress, which can assist students in their learning introspection. In short, ITSs can help students avoid cognitive impasses, errors, and misconceptions, and can also alleviate negative affect. They use process supervision to promote students' reflection and improve their cognition and metacognition.

### 3. Research design

This paper seeks to identify the affective experiences and effectiveness of using an ITS compared to a non-ITS learning environment. The learning content covered the ecological relationship between wolves, hunters, cows, deer, grass, rainfall, and other concepts about a forest ecosystem. For example, some lessons present these concepts in terms of an increase or decrease in water and food availability and how this affects the animal population.

The experimental group used Betty's Brain, an ITS developed by a combination of computer science, psychology, and education researchers in the engineering school of Vanderbilt University. The system uses virtual teachers (Mr. Davis) and virtual students (Betty) to intervene and guide students' cognition and affect. The ITS consists of a "Causal Map," "Science Book," "Notes," "Quiz Results," and "Teacher's Guide." The control class used a non-ITS (F\_S), which is based on Moodle 2.8 and covers the same domain and content as Betty's Brain, including "Science Book," "Notes," "Quiz Results," and "Teacher's Guide." F\_S does not have virtual teachers and students, and participants used Microsoft Word to build causal relationships.

Participants included 72 students in the seventh grade of a middle school in Changchun. Participants consisted of 35 boys and 37 girls. All the students had no experience with using ITSs, and the two classes were taught by the same teachers.

The experimental process was mainly divided into a teaching stage, autonomous learning stage and an after-class interview. In the teaching stage, teachers guided students through content with the theme of "forest ecosystem"

and taught them how to use a “Causal Map,” “Science Book,” “Notes,” “Quiz Results,” and “Teacher’s Guide” in 3 minutes. In the autonomous learning stage, students needed to construct a causality diagram in 25 minutes, during which we collected the video data. Afterwards some students needed to complete interviews lasting no longer than 13 minutes.

This study used 46 Mosheng RQES008 HD digital cameras to capture facial expressions with a USB 2.0 interface, and the cameras were assembled and installed on every computer. The coding of the types of affect was based on previous coding schemes used by McDaniel et al. (2007) and Altuwairqi et al. (2021). We referred to the facial expressions in the video data to judge students’ affect. Each coding result was recorded in a table, like Figure 2.

Time	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Stu1	Confusion	Boredom	Confusion	Confusion	Confusion	Boredom	Confusion	Confusion	Confusion	Confusion
Stu2	Confusion	None	Confusion	Flow	Flow	Flow	Flow	Confusion	Confusion	Confusion
Stu3	Confusion	Flow	Confusion	Flow	Confusion	Flow	Flow	Confusion	Boredom	Confusion
Stu4	Boredom	Boredom	Confusion	Flow	Flow	Flow	Confusion	Flow	Flow	Confusion

Figure 2. Affect encoding sample in the control group

As is shown in Figure 2, the first row represents time, which is used to mark 25 encodings. The first column represents students’ identity, for example, stu1 as the first student. Affect for each timepoint were coded as either boredom, flow, confusion, frustration, surprise, delight, or none. After further observation and discussion, because some affective states, such as frustration or surprise were very rare, we only considered “boredom,” “flow,” “confusion,” “delight” and “no affect” in this paper.

## 4. Results and analysis

Two sets of 25-minute videos during the autonomous learning stage were used for analysis, which coded the following states of: “boredom,” “flow,” “confusion,” “delight” and “no affect.” Our coding process was handled and reviewed by two experimenters in charge. If the number of the matching codes is  $x$ , and the number of codes for each person is  $y$ , then the quotient ( $x/y$ ) can be defined as the coding consistency. The two experimenters simultaneously encoded two of the same samples in order. After comparing and contrasting between both results, there were 37 matches in the 50 coded data points. In short, these were recorded every 30 seconds during the autonomous learning stage, and the video coding consistency between the two raters is (37/50) 74%.

### 4.1. Affective cumulative analysis

To analyze the overall affect distribution, the affective states of each group are summarized below. Taking the autonomous learning stage into account, twenty-five minutes of activities were recorded every 30 seconds. The numbers in the first row represent 50 different recordings, and the data represents the frequency of the corresponding affect. Taking the 21<sup>st</sup> encoding in the 22<sup>nd</sup> column in the control group as an example, 1 student showed “boredom,” 11 were in “flow,” 6 were “confused,” 2 showed “delight,” and for 14 of the students we were unable to determine their affect because they were off camera. Accordingly, the cumulative frequency of affect is summarized in Tables 1 and 2.

Table 1. The affective accumulation table of the control group

Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Boredom	1	2	1	1	0	1	1	0	1	1	3	1	4	2	2	2	1	2	2	0	1	0	2	2	1
Flow	10	10	9	13	11	13	11	11	13	13	12	15	13	12	11	14	10	14	14	16	11	13	11	10	9
Confusion	6	4	8	4	7	2	6	7	3	6	4	2	3	6	4	3	7	1	1	2	6	3	5	7	3
Delight	4	1	1	1	1	1	0	0	1	1	1	1	1	1	2	0	2	1	2	0	2	0	0	2	0
None	13	17	15	15	15	17	16	16	16	13	14	15	13	13	15	15	14	16	15	16	14	18	16	13	21
Number	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Boredom	1	0	1	0	1	3	1	2	0	0	0	2	1	1	0	0	0	0	0	0	1	1	1	0	2
Flow	9	12	9	9	11	10	8	6	7	9	7	6	9	8	7	8	7	9	5	6	7	5	7	7	4
Confusion	6	3	4	3	6	3	4	7	7	7	6	3	5	1	3	4	4	1	2	1	0	0	0	0	0
Delight	2	3	2	1	1	0	2	2	2	2	2	1	1	2	1	0	0	2	1	1	0	0	0	0	0
None	16	16	18	21	15	18	19	17	18	16	19	22	18	22	23	22	23	22	26	26	26	28	26	27	28

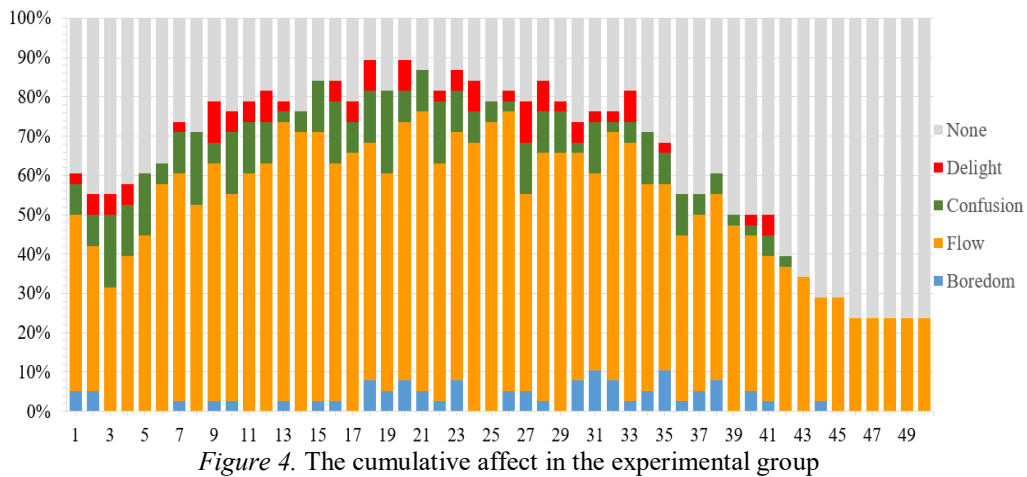
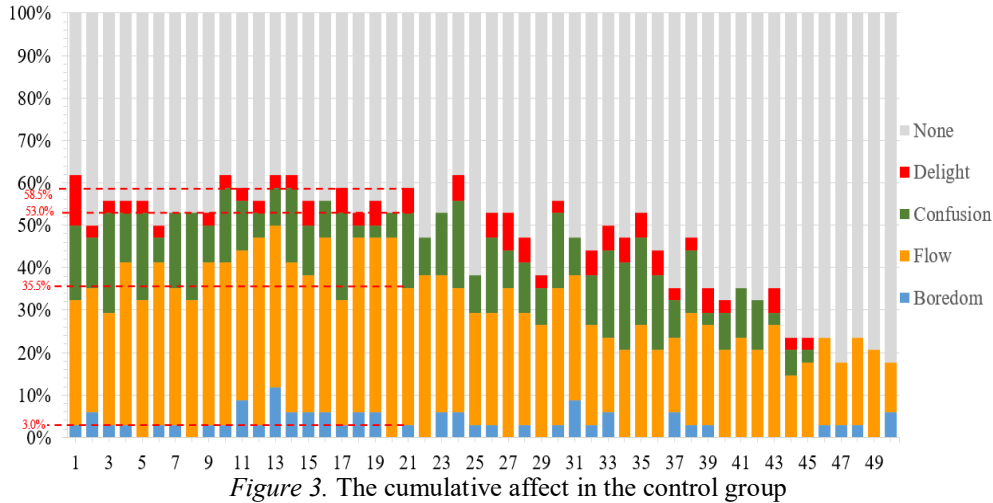
Table 2. The affective accumulation table of the experimental group

Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Boredom	2	2	0	0	0	0	1	0	1	1	0	0	1	0	1	1	0	3	2	3	2	1	3	0	0
Flow	17	14	12	15	17	22	22	20	23	20	23	24	27	27	26	23	25	23	21	25	27	23	24	26	28
Confusion	3	3	7	5	6	2	4	7	2	6	5	4	1	2	5	6	3	5	8	3	4	6	4	3	2
Delight	1	2	2	2	0	0	1	0	4	2	2	3	1	0	0	2	2	3	0	3	0	1	2	3	0
None	15	17	17	16	15	14	10	11	8	9	8	7	8	9	6	6	8	4	7	4	5	7	5	6	8
Number	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Boredom	2	2	1	0	3	4	3	1	2	4	1	2	3	0	2	1	0	0	1	0	0	0	0	0	0
Flow	27	19	24	25	22	19	24	25	20	18	16	17	18	18	15	14	14	13	10	11	9	9	9	9	9
Confusion	1	5	4	4	1	5	1	2	5	3	4	2	2	1	1	2	1	0	0	0	0	0	0	0	0
Delight	1	4	3	1	2	1	1	3	0	1	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0
None	7	8	6	8	10	9	9	7	11	12	17	17	15	19	19	19	23	25	27	27	29	29	29	29	29

According to the accumulated frequency (as shown in Tables 1 and 2), we observed the following:

- “Flow” and “none” frequently occurred during their learning processes followed with “confusion” in each group, “boredom” and “delight” occurred the least, according to the accumulated frequency.
- Affect changes over time, so it follows that each affect here fluctuates across the recordings. For example, the frequency of “flow” is 9 to 16 in the first 25 recordings in the control group, and “flow” is 4 to 12 in the last 25 recordings.
- Many of the affective states in the groups were coded as “none.” This is because the task was not completed, and the camera was disconnected. In this case, it is impossible to determine some affective states. For instance, the frequency of “none” fluctuated from 13 to 21 in the first 25 recordings, and then the frequency increased from 15 to 28 in the last 25 recordings in the control group.

To summarize the overall affect distribution of each group, the cumulative data is represented by a bar graph as shown in Figures 3 and 4.



Twenty-five minutes of activity were recorded every 30 seconds, and 50 recordings are shown on the x-axis. The y-axis represents the cumulative percentage of each affect. Taking the 21<sup>st</sup> recording in the x-axis of the control group as an example, 3.0% of students were “bored,” 32.5% were in “flow,” 17.5% were “confused,” 5.5% showed “delight,” and 41.5% could not be determined.

Figures 3 and 4 display the proportion of each affect at 50 different video captures throughout the learning sessions. For example, the proportion of “flow” in the experimental group is higher than that of the control group throughout the entire session. In the control group, “flow” increased from time 1 to time 3, peaked at the 20<sup>th</sup> recording (about 47%), and then slowly declined to 11% by the 50<sup>th</sup> recording. Comparatively, “flow” in the experimental group fluctuated from approximately 32% to around 45% from time 1 to time 5, then gradually increased to about 74% at the 25<sup>th</sup> recording. It then gradually declined to roughly 24% by the 50<sup>th</sup> recording.

In general, there was less “confusion” in the experimental group than the control group. In the control group, the proportion of “confusion” started at 17% and increased to about 21% by the 5<sup>th</sup> recording. From the 6<sup>th</sup> to the 45<sup>th</sup> recording, “confusion” fluctuated between approximately 3% to roughly 21%, and afterwards students did not show “confusion.” In the experimental group, the proportion of “confusion” fluctuated between about 2% to around 21% up to the 37<sup>th</sup> recordings, and students did not display “confusion” from the 43<sup>rd</sup> to the 50<sup>th</sup> recording.

In summary, both the cumulative frequency and percentage of “flow” was significantly higher in the experimental group than that in the control group. The cumulative percentage of “confusion” was higher in the control group than that in the experimental group. There was no significant difference between the two conditions for the other affective states.

## 4.2. Analysis of the difference in each group

Generalized Sequential Querier (GSEQ) can be used to analyze sequence observation data. GSEQ can be used to perform coding and output the frequency of affect transitions (see Table 3). Here, the data consists of the frequency of transitions from the  $i^{\text{th}}$  affect to the  $j^{\text{th}}$  affect, denoted as  $X_{ij}$ . The variable  $i$  represents the affect index of columns,  $j$  represents the affect index of rows,  $N$  represents the type of affect coded, and the range of changes both  $i$  and  $j$  is  $[1, N]$ .

Table 3. Joint frequency table

	Given	Boredom	Flow	Confusion	Delight	None	Totals
Control group	Boredom	13	14	10	4	10	51
	Flow	11	341	59	17	61	489
	Confusion	10	56	74	10	39	189
	Delight	3	18	7	10	15	53
	None	15	54	33	8	774	884
	Totals	52	483	183	49	899	1666
	Given	Boredom	Flow	Confusion	Delight	None	Totals
Experimental group	Boredom	7	21	14	3	11	56
	Flow	21	795	69	29	49	963
	Confusion	11	71	49	7	16	154
	Delight	4	27	10	9	8	58
	None	11	41	9	9	561	631
	Totals	54	955	151	57	645	1862

The frequency of each type of affective transition is different from each group (see Table 3). A total of 1666 transformations were observed in the control group and 1862 changes in the experimental group. Frequent patterns (frequency greater than or equal to 30) include: “flow/confusion/none  $\rightarrow$  flow/confusion/none” in the control group, “flow/confusion/none  $\rightarrow$  flow,” “flow/confusion  $\rightarrow$  confusion” and “none/flow  $\rightarrow$  none” in the experimental group. Some frequent transition patterns we observed in both conditions are: “flow/confusion/none  $\rightarrow$  flow,” “flow/confusion  $\rightarrow$  confusion,” and “none/flow  $\rightarrow$  none.” The transition patterns of “confusion  $\rightarrow$  none,” and “none  $\rightarrow$  confusion” were frequent only in the control group. The frequencies of “flow/confusion/delight  $\rightarrow$  bored,” “bored/flow/confusion/delight  $\rightarrow$  flow,” “bored/flow/delight  $\rightarrow$  confused,” “flow/none  $\rightarrow$  delight,” and “bored  $\rightarrow$  no affect” were all significantly higher in the experimental group than in the control group. The affect transition frequencies of “bored/none  $\rightarrow$  bored,” “none  $\rightarrow$  flow,”

“*confusion/none* → *none/confused*,” “*bored/confused/delight* → *delight*,” and “*flow/delight* → *none*” were significantly higher in the control group than in experimental group.

The GSEQ tool calculated the expected frequency of affect transitions shown in Table 4, by using the observed frequencies shown in Table 3 and the  $M_{ij}$  formula.

$$M_{ij} = \frac{(X_{i+}) * (X_{+j})}{\sum_{i=1}^N \sum_{j=1}^N X_{ij}} = \frac{(\sum_{j=1}^N X_{ij}) * (\sum_{i=1}^N X_{ij})}{\sum_{i=1}^N \sum_{j=1}^N X_{ij}} = \frac{X_{i,j=totals} * X_{i=totals,j}}{X_{i=totals,j=totals}} \quad (1)$$

Table 4. Expected frequency table

	Given	Boredom	Flow	Confusion	Delight	None
Control group	Boredom	1.592	14.786	5.602	1.500	27.520
	Flow	15.263	141.769	53.714	14.382	263.872
	Confusion	5.899	54.794	20.761	5.559	101.987
	Delight	1.654	15.366	5.822	1.559	28.600
	None	27.592	256.286	97.102	26.000	477.020
Experimental group	Given	Boredom	Flow	Confusion	Delight	None
	Boredom	1.624	28.722	4.541	1.714	19.398
	Flow	27.928	493.912	78.095	29.480	333.585
	Confusion	4.466	78.985	<b>12.489</b>	<b>4.714</b>	<b>53.346</b>
	Delight	1.682	29.748	4.704	1.776	<b>20.091</b>
	None	<b>18.300</b>	323.633	<b>51.171</b>	19.316	<b>218.579</b>

Expected frequency  $M_{ij}$  refers to the product of  $X_{i,j=totals}$  (sum of the frequencies at which all affective states turn into the  $j^{\text{th}}$  affect) multiplied by  $X_{i=totals,j}$  (sum of the frequencies at which the  $i^{\text{th}}$  affective state turns into all affective states) and the quotient of  $X_{i=totals,j=totals}$  (sum of all affect transitions). In other words, this formula is used to calculate the expectation of each transition, which is placed in all the transition processes. The expected affect frequency is different from the initial frequency, such as the joint frequency of “*flow*” to “*confusion*” equals 59, and the expected frequency of the transition of “*flow*” to “*confusion*” equals 53.714.

Some of the expected frequencies of affect transitions were significantly different between the two groups. The frequencies of “*bored/flow/delight* → *bored*,” “*bored/confusion/delight/none/flow* → *flow*,” “*flow* → *confused*,” “*bored/flow/delight* → *delight*” and “*none* → *none*” in the experimental group are higher than those in the control group. The frequencies of “*confused/none* → *bored*,” “*bored/confused/delight/none* → *confused*,” “*confused/none* → *delight*,” and “*bored/confused/delight/none* → *none*” in the control group are higher than those in the experimental group.

Table 5. Summary table of adjusted residuals of affective transformation

	Given	Boredom	Flow	Confusion	Delight	None
Control group	Boredom	<b>9.330</b>	-0.246	<b>2.000</b>	<b>2.104</b>	<b>-4.999</b>
	Flow	-1.319	<b>23.624</b>	0.910	0.834	<b>-21.899</b>
	Confusion	1.822	0.205	<b>13.153</b>	<b>2.031</b>	<b>-9.763</b>
	Delight	1.080	0.811	0.526	<b>6.974</b>	<b>-3.809</b>
	None	<b>-3.555</b>	<b>-21.887</b>	<b>-10.064</b>	<b>-5.230</b>	<b>29.250</b>
Experimental group	Given	Boredom	Flow	Confusion	Delight	None
	Boredom	<b>4.347</b>	<b>-2.096</b>	<b>4.701</b>	1.013	<b>-2.395</b>
	Flow	-1.915	<b>27.936</b>	-1.545	-0.129	<b>-27.737</b>
	Confusion	<b>3.276</b>	-1.344	<b>11.253</b>	1.116	<b>-6.604</b>
	Delight	1.843	-0.733	<b>2.588</b>	<b>5.595</b>	<b>-3.390</b>
	None	<b>-2.130</b>	<b>-27.685</b>	<b>-7.564</b>	<b>-2.932</b>	<b>35.234</b>

Note.  $|Z_{ij}| > 1.96$ .

This paper uses the  $Z_{ij}$  formula to calculate the adjusted residual value given the data shown in Table 4 and the joint frequencies in Table 5.

$$Z_{ij} = \frac{X_{ij} - M_{ij}}{\sqrt{M_{ij} * \left(1 - \frac{M_{ij}}{X_{+j}}\right) * \left(1 - \frac{M_{ij}}{X_{i+}}\right)}} = \frac{X_{ij} - M_{ij}}{\sqrt{M_{ij} * \left(1 - \frac{M_{ij}}{\sum_{i=1}^N X_{ij}}\right) * \left(1 - \frac{M_{ij}}{\sum_{j=1}^N X_{ij}}\right)}} = \frac{X_{ij} - M_{ij}}{\sqrt{M_{ij} * \left(1 - \frac{M_{ij}}{X_{i=totals,j}}\right) * \left(1 - \frac{M_{ij}}{X_{i,j=totals}}\right)}} \quad (2)$$

$Z_{ij}$  is used to calculate the difference between the observation and the expectation. We use the formula (Haberman, 1979) to execute and compute the adjusted residual value. The product of probability of neither belonging to  $X_{i,j=total,s}$  nor belonging to  $X_{i=total,j}$  is used as the weight of  $M_{ij}$ . The difference between the actual value and the expected value is used as the dividend, and the root of the expected value including weight is used as the divisor. The quotient of the two is called the adjusted residual value. The adjusted residuals are similar to Z-scores;  $Z_{ij}$  is normally distributed, and the Z-test can be used to test the statistical significance. According to the standard normal distribution in the Z-value table,  $Z_{ij}$  is substituted into the normal distribution to find the corresponding probability  $P$ -value. Also,  $|Z_{ij}| > 1.96$  (95% confidence interval) is selected to indicate a significant change in affect, which is marked in bold font.

According to Table 5, the significant  $|Z_{ij}|$  is marked on the affect conversion graph, and the arrows point to the next affect of the transition, and thicker lines indicate more significance of the affect transitions. The conversion relationship is drawn, as shown in Figures 5 and 6.

We observed the following affective conversions.

- There are repeating or recurrent patterns of affect conversions, which include: “boredom→boredom/confusion/none,” “flow→flow/none,” “confusion→confusion/none,” “delight→delight/none,” and “none→boredom/flow/confusion/delight/none.” The “boredom→boredom,” “confusion→confusion,” “confusion→none,” “none→confusion,” “delight→delight,” “delight→none,” “none→delight,” “boredom→none,” “none→boredom,” and “confused→delight” transitions are more significant than rest of the affect transition in both the control group and experimental group.
- There are different transition patterns between the two groups. For example, “boredom/confusion→delight” is significant in the control group, but not in the experimental group. “Delight→confusion” and “confusion→boredom” are significant patterns in the experimental group, but not in the control group. Additionally, the transition of “boredom→delight” and “confused→delight” is significant in the control group, but not the experimental group. The transition of “delight→confused” and “confused→boredom” are significant in the experimental group, but not the control group.

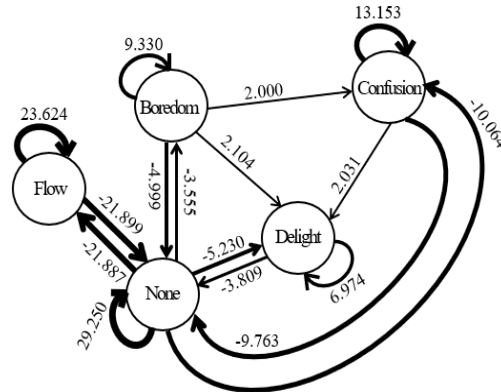


Figure 5. Affective conversion in the control group

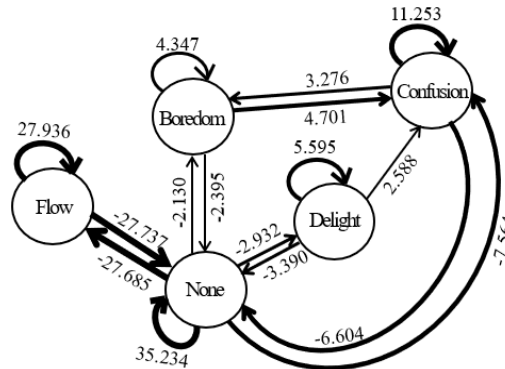


Figure 6. Affective conversion in the experimental group

In summary, based on the data analysis above, we can see the specific difference and significance in each transition in each group. However, we cannot compare the specific Z-value across different groups, and we can only compare the Z-value in the inner group.

### 4.3. Analysis of the difference between the two groups of affect

This paper analyzes the significance of the difference between the two groups, by using a multinomial processing tree (MPT) to analyze the frequency of transition (see the Table 3) in a general processing tree (GPT) software. The results are shown in Table 6.

In Table 6, the parameters (PA) in the first column represents every transition in each group, for example, the CAA pattern refers to the  $A \rightarrow A$  in the control group, and EAA refers to the  $A \rightarrow A$  in the experimental group. The star sign (\*) in the first column in the table refers to the parameter(s) that are restricted as constant. Additionally, in the second column, EV is the estimated value of the parameter. SD refers to standard deviation in the third column. For the confidence intervals (CI), in the fourth column, LO represents the lower limit of the confidence interval and UP represents the upper limit of the confidence interval in the fifth column. “Sig of Difference” refers to the statistical significance of the difference in artificial processing. “LO\_SI” shows the lower limit of the significance of the difference, and the last column shows the upper limit of the significance of the difference (UP\_SI).

By using the confidence interval of the parameter estimation values in the groups, we can directly assess the difference of the significance of the difference of the parameter. The LO\_SI of difference value equals the upper limit of the confidence interval in the experimental group. The negative sign corresponds to the lower limit of the confidence interval in the control group. Similarly, the UP\_SI of difference value equals the upper limit of the confidence interval in the control group minus corresponds to the lower limit of the confidence interval in the experimental group. The difference is significant if the one of the LO\_SI value and UP\_SI value is less than 0, otherwise it cannot be intuitively judged.

Table 6. Analysis in model multinomial processing tree

Control Group					Experimental Group					Sig of Difference	
PA	EV	SD	95% CI		PA	EV	SD	95% CI			
			LO	UP				LO	UP	LO	SI
CA*	0.031	constant			EA*	0.030	constant				
CAA	0.255	0.061	0.135	0.375	EAA	0.125	0.044	0.038	0.212	0.076	0.336
CAB	0.275	0.062	0.152	0.397	EAB	0.375	0.065	0.248	0.502	0.350	0.149
CAC	0.196	0.056	0.087	0.305	EAC	0.250	0.058	0.137	0.363	0.276	0.168
CAH	0.078	0.038	0.005	0.152	EAH	0.054	0.030	-0.005	0.113	0.108	0.158
CB*	0.294	constant			EB*	0.517	constant				
CBA	0.022	0.007	0.009	0.036	EBA	0.022	0.005	0.013	0.031	0.022	0.023
<b>CBB</b>	<b>0.697</b>	<b>0.021</b>	<b>0.657</b>	<b>0.738</b>	<b>EBB</b>	<b>0.826</b>	<b>0.012</b>	<b>0.802</b>	<b>0.850</b>	<b>0.193</b>	<b>-0.064</b>
<b>CBC</b>	<b>0.121</b>	<b>0.015</b>	<b>0.092</b>	<b>0.150</b>	<b>EBC</b>	<b>0.072</b>	<b>0.008</b>	<b>0.055</b>	<b>0.088</b>	<b>-0.004</b>	<b>0.094</b>
CBH	0.035	0.008	0.019	0.051	EBH	0.030	0.006	0.019	0.041	0.022	0.032
CC*	0.113	constant			EC*	0.083	constant				
CCA	0.053	0.016	0.021	0.085	ECA	0.071	0.021	0.031	0.112	0.091	0.054
<b>CCB</b>	<b>0.296</b>	<b>0.033</b>	<b>0.231</b>	<b>0.361</b>	<b>ECB</b>	<b>0.461</b>	<b>0.040</b>	<b>0.382</b>	<b>0.540</b>	<b>0.309</b>	<b>-0.021</b>
CCC	0.392	0.036	0.322	0.461	ECC	0.318	0.038	0.245	0.392	0.070	0.216
CCH	0.053	0.016	0.021	0.085	ECH	0.045	0.017	0.013	0.078	0.057	0.072
CG*	0.472	constant									
CGA	0.017	0.004	0.008	0.025	EGA	0.017	0.005	0.007	0.028	0.019	0.018
CGB	0.061	0.008	0.045	0.077	EGB	0.065	0.010	0.046	0.084	0.039	0.031
<b>CGC</b>	<b>0.037</b>	<b>0.006</b>	<b>0.025</b>	<b>0.050</b>	<b>EGC</b>	<b>0.014</b>	<b>0.005</b>	<b>0.005</b>	<b>0.024</b>	<b>-0.001</b>	<b>0.045</b>
CGH	0.009	0.003	0.003	0.015	EGH	0.014	0.005	0.005	0.024	0.021	0.010
CH*	0.032	constant			EH*	0.031	constant				
CHA	0.057	0.032	-0.006	0.119	EHA	0.069	0.033	0.004	0.134	0.140	0.115
CHB	0.340	0.065	0.212	0.467	EHB	0.466	0.066	0.337	0.594	0.382	0.130
CHC	0.132	0.047	0.041	0.223	EHC	0.172	0.050	0.075	0.270	0.229	0.148
CHH	0.189	0.054	0.083	0.294	EHH	0.155	0.048	0.062	0.248	0.165	0.232

Note. A refers to boredom, B refers to flow, C refers to confusion, H refers to delighted, G refers to no affect.

According to the MPT model (as shown in Table 6), some affect transitions are significantly different between the two groups, and the significant transitions are bolded in the table. For instance, the  $B \rightarrow B$  transition in the experimental group (estimated value = 0.826,  $SD$  = 0.012, lower limit of the confidence interval = 0.802) is significantly higher than the  $B \rightarrow B$  in the control group (estimated value = 0.697,  $SD$  = 0.021, upper limit of confidence interval = 0.738); the  $C \rightarrow B$  in the experimental group (estimated value = 0.461,  $SD$  = 0.040, lower



limit of confidence interval= 0.382) is significantly higher than the C→B in control group (estimated value = 0.296, SD = 0.033, upper limit of confidence interval = 0.361). The B→C (estimated value = 0.072, SD = 0.008, upper confidence interval= 0.088) in the experimental group is significantly smaller than B→C in the control group (estimated value = 0.121, SD = 0.015, lower confidence interval = 0.092). The G→C in the experimental group (estimated value = 0.014, SD = 0.005, upper confidence interval = 0.024) is significantly smaller than in the control group (estimated value = 0.037, SD = 0.006, lower confidence interval = 0.025). The results of a chi-square test indicate that there are significant differences in the affective transitions between both groups ( $\chi^2[9] = 0.01988$ ).

According to the affect transitions of the two groups in Table 3, this paper calculates the difference between combined frequency of affect changes in two groups, denoted as  $X'_{ij}$ , which represents the difference of frequency between the  $i^{th}$  affect and the  $j^{th}$  affect transition, see Table 7.

Table 7. The combined frequency of the difference between the two groups

Given	Boredom	Flow	Confusion	Delight	None	Totals
Boredom	-6	7	4	-1	1	5
Flow	10	454	10	12	-12	474
Confusion	1	15	-25	-3	-23	-35
Delight	1	9	3	-1	-7	5
None	-4	-13	-24	1	-213	-253
Totals	2	472	-32	8	-254	196

According to the value of  $X'_{ij}$ , the combined frequency is different from the initial joint frequency. Some transitions in the experimental group were less than those in the control group, such as, “none→none,” “confusion→confusion,” “none→confusion,” “confusion→none,” and “none→flow.” Some affective changes in the experimental group are more frequent than those in the control group, for instance, “flow→flow,” “confusion→flow,” and “flow→delight.”

According to the frequency of transformation in Table 7, this article calculates the expected transformation of the difference between the experimental group and the control group, denoted as  $M'_{ij}$ , which represents the expected frequency of transformation from the  $i^{th}$  affect to the  $j^{th}$  affect.

$$M'_{ij} = \frac{(X'_{i+}) * (X'_{+j})}{\sum_{i=1}^I \sum_{j=1}^J X'_{ij}} = \frac{(\sum_{j=1}^J X'_{ij}) * (\sum_{i=1}^I X'_{ij})}{\sum_{i=1}^I \sum_{j=1}^J X'_{ij}} = \frac{X'_{i,j=totals} * X'_{i=totals,j}}{X'_{i=totals,j=totals}} \quad (3)$$

Table 8. The expected frequency of the difference between the two groups

Given	Boredom	Flow	Confusion	Delight	None
Boredom	<b>0.051</b>	<b>12.041</b>	-0.816	<b>0.204</b>	-6.480
Flow	<b>4.837</b>	<b>1141.469</b>	-77.388	<b>19.347</b>	-614.265
Confusion	-0.357	-84.286	<b>5.714</b>	-1.429	<b>45.357</b>
Delight	<b>0.051</b>	<b>12.041</b>	-0.816	<b>0.204</b>	-6.480
None	-2.582	-609.265	<b>41.306</b>	-10.327	<b>327.867</b>

According to the expected frequency (as shown in Table 8), the difference frequency between the two groups is also different from the combined frequency (as shown in Table 7). Some transitions in the experimental group are less than those in the control group including: “flow→none,” “none→flow,” “confusion→flow,” “flow→confusion,” “none→delight,” “boredom→confusion,” “boredom→none,” “delight→none,” “none→boredom,” “confusion→delight,” “delight→confusion,” and “confusion→boredom.” Some changes in the experimental group are more likely than that in the control group including: “flow→flow,” “none→none,” “confusion→none,” “none→confusion,” “flow→delight,” “delight→flow,” “boredom→flow,” “confusion→confusion,” “flow→boredom,” “boredom→delight,” “delight→delight,” “delight→boredom,” and “boredom→boredom.”

$Z'_{ij}$  is used to calculate the difference between observations and expectations. The product of the probabilities of neither belonging  $X'_{i,j=totals}$  nor belonging  $X'_{i=totals,j}$  is used as the weight of  $M'_{ij}$ . The difference between the initial value and the expected value is used as a dividend, and weighted expected value is used as a divisor, and a quotient of the two is called the adjusted residual value.

According to frequency (Table 7) and expectation (Table 8), this paper calculates the adjusted residual value, which is expressed as  $Z'_{ij}$  (see Table 9).

$$Z'_{ij} = \frac{x'_{ij} - M'_{ij}}{\sqrt{M'_{ij} \left(1 - \frac{M'_{ij}}{x'_{+j}}\right) \left(1 - \frac{M'_{ij}}{x'_{i+}}\right)}} = \frac{x'_{ij} - M'_{ij}}{\sqrt{M'_{ij} \left(1 - \frac{M'_{ij}}{\sum_{i=1}^N x'_{ij}}\right) \left(1 - \frac{M'_{ij}}{\sum_{j=1}^N x'_{ij}}\right)}} = \frac{x'_{ij} - M'_{ij}}{\sqrt{M'_{ij} \left(1 - \frac{M'_{ij}}{x'_{i=totals,j}}\right) \left(1 - \frac{M'_{ij}}{x'_{i,j=totals}}\right)}} \quad (4)$$

Table 9. The residual of the difference between the two groups

Given	Boredom	Flow	Confusion	Delight	None
Boredom	-27.277	-1.240	<b>5.007</b>	-2.757	<b>1.964</b>
Flow	<b>1.981</b>	<b>-14.398</b>	<b>7.734</b>	-1.432	<b>13.466</b>
Confusion	<b>2.103</b>	<b>8.395</b>	<b>-10.973</b>	-1.237	<b>-6.170</b>
Delight	<b>4.278</b>	-0.748	<b>3.967</b>	-2.757	-0.137
None	-0.586	<b>13.450</b>	<b>-6.225</b>	<b>2.378</b>	<b>-13.025</b>

Note.  $|Z'_{ij}| > 1.96$ .

As shown in Table 9, the significant  $Z'_{ij}$  is marked on the affect conversion graph, with the arrows pointing to the next affect of the transition. Thicker lines indicate more significance of the affect transitions. The conversion relationship is drawn, as shown in Figure 7.

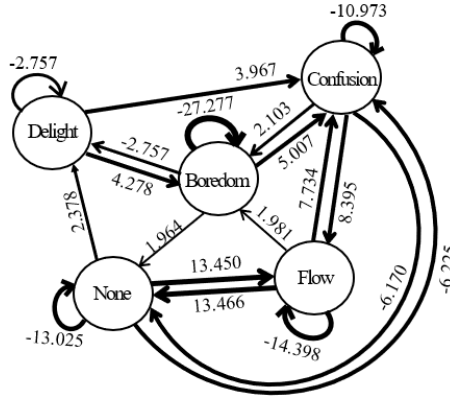


Figure 7. Affective conversion diagram of the groups

According to the residuals in Figure 7, some patterns with significant differences between the groups were observed, which include: “none→flow,” “flow→none,” “confusion→flow,” “flow→confusion,” “boredom→confusion,” “boredom→boredom,” “flow→flow,” “confused→confused,” and “none→none.”

In summary, there are significant differences between the two groups in affective transitions, especially when negative affect transformed into positive affect, such as “confusion/none→flow” and “none/boredom→delight.” There are significant differences in negative affect changes. Positive affect transformed into positive affect, such as “flow→flow” and “delight→delight.” Additionally, there are positive changes into negative, such as “flow/delight→bored/confused.” There are also negative affect changes into negative, such as “confusion→boredom” and “boredom→confusion.” Finally, there are also significant differences for the transitions of “boredom/flow/confusion/none→none.”

## 5. Conclusion

This paper first summarized the affect regulation methods supported by teachable agents and the affect regulation processes in ITSs. Four ITS functions that can be used to detect or help regulate affect were described. To supervise and adjust negative affect, ITSs use intelligent algorithms and technologies to analyze learning data (e.g., cognition and mood), determine learning affect, and then provide reasonable and flexible strategies (e.g., refined learning materials) or rigid strategies (e.g., simple rehearsing).

According to the accumulation analysis, students in the experimental group were prone to “confusion” and “boredom,” but they spent more time in a “flow” state. It should be noted that while “confusion” is typically thought of a negative affective state, research has shown that it can be beneficial to learning when it does not

lead to “*frustration*,” “*boredom*,” and disengagement (D’Mello, et al., 2014). With the system’s help, students adjusted these negative affective states to “*delight*” and “*flow*.” The use of scaffolds (such as prompts, tests, responses, and notes) often showed that students were surprised about their results. The “*flow*” state was more common in the experimental group than the control group, which suggests a higher degree of concentration in the experimental group. Therefore, the affect of the experimental group was more positive than in the control group.

Lag sequence analysis was used to analyze the different affect transitions in each group (see Table 3). The quantity of the affect transitions in the two groups is different and the frequency of standardized emotional transitions of two groups are also different (Table 4). The adjusted residual value of each affect transition in each group is standardized and the size indicates differing scales of affect transition in the groups. Significant differences were observed for each type of affect transition.

To further explore the differences in emotional transitions between the two groups, we used the MPT method to analyze the differences in affect transitions between the two groups. The results revealed significant differences in the transitions of the two groups in individual affect transition types. For example, the transitions from “*flow*” to “*flow*” (i.e., staying in a “*flow*” state) and “*confusion*” to “*flow*” (i.e., resolving some “*confusion*”) in the experimental group are significantly higher than the same transitions in the control group. The transition from “*flow*” to “*confusion*” (i.e., reach an impasse) and “*none*” to “*confusion*” in the experimental group are significantly lower than the same in the control group. There are not only internal differences in each group, but also significant differences between the two groups, which we observed from our lag sequence analysis (see Tables 7 and 9). For example, in the control condition, the likelihood of students remaining in a “*bored*” state (“*bored* → *bored*”) is stronger (i.e., more significant) than in the experimental group. Comparatively, students in the experimental condition remained in a confused state less frequently than in the control condition.

We suggest two reasons for the occurrence of positive affect regulation, based on our observations of the video data and after-class interviews. First, learning with an ITS is engaging and has game-like features. With Betty’s Brain, students learned about biological relationships and exercised thinking strategies to solve a task. A second reason why positive affect regulation occurred may be in part due to the virtual characters that help students find content related to the task at hand. This is consistent with the experimental conclusion of Segedy et al. (2014), who also used Betty’s Brain for their study. They found that the ITS provides students with the necessary support in a timely manner, so that students can apply cognitive and metacognitive strategies to solve “cause-and-effect” problems. They also concluded that the system helps promote students’ deep learning and guides them to use suitable strategies to solve problems. Some students in the experimental group began to use more optimized logic to complete tasks. This indicates that they consciously took advantage of the system’s cognitive and metacognitive scaffolds to assess causality. These help students better regulate their affect and enhance their learning effectiveness. As one participant said, “I study causality very seriously. I always hope to teach students the correct knowledge. Therefore, I am confident that I can complete this task.” This sentiment is consistent with the Kobylińska and Karwowska (2015) research on using automated affect regulation to influence students’ negative affective experience. Students in the experimental condition appeared to be attending to the tasks at hand, given their time spent on task and mouse click frequency within the interface. While students were attending to the tasks, the ITS helped students become aware of their affect through dialogue. Students could then report their affective state to the teacher agent. Accurately grasping affect perception, evaluation, and expression requires understanding affect and affective knowledge, controlling affect and affective intelligence development, and thereby enhancing students’ ability to recognize, regulate and manage affect.

This study took into account the affect transitions both within and between the two groups. However, this study has some limitations. Differences in the learning level and cognitive development of students varies by region, so the conclusions of this experiment may not be universal. It is also necessary to combine iterative experiments to avoid time shortage and contingency problems. It is possible to gain more fruitful results after multiple rounds of repeated experiments. We encoded students’ affect by human observation, so the results may have some bias, so it is necessary to adopt artificial intelligent technologies to analyze the specific information automatically. A large amount of data is needed to further explore the deep relationship between affect regulation and cognition. Likewise, future studies can capture video and audio data more frequently, which would strengthen the reliability of the results. It is also necessary to collect both cognitive and metacognitive data at the same time. This allows for an exploration of the in-depth relationship between affect and metacognition by using a multi-branch tree analysis. Hwang et al. (2020) reported that changes in affect result in performance changes. Therefore, future studies of affect regulation in ITSs would benefit from tracking both affect changes and performance changes over time.

## Acknowledgement

This research was partially sponsored by the National Science Foundation under the award The Learner Data Institute (award #1934745). The opinions, findings, and results are solely the authors' and do not reflect those of the funding agencies. This study was also supported by National Social Science Foundation of China for education programs (Project Approval # BCA190076). This study was also supported by the National Natural Science Foundation of China (Project Approval # 62077012).

## References

- Altuwairqi, K., Jarraya, S. K., Allinjaw, A., & Hammami, M. (2021). A New emotion-based affective model to detect student's engagement. *Journal of King Saud University - Computer and Information Sciences*, 33(1), 99-109.
- Alexandra, J. M., Ocumpaugh, J., Baker, R. S., Slater, S., Paquette, L., Jiang, Y., Karumbaiah, S., Bosch, N., Munshi, A., Moore, A., & Biswas, G. (2019). Affect sequences and learning in Betty's Brain. In D. Azcona & R. Chung (Eds.), *Proceedings of 9th International Conference on Learning Analytics & Knowledge (LAK 19)* (pp. 383-390). New York, NY: Association for Computing Machinery. doi:10.1145/3303772.3303807
- Alkhalaf, A. M. (2018). Positive and negative affect, anxiety, and academic achievement among medical students in Saudi Arabia. *International Journal of Emergency Mental Health and Human Resilience*, 20(2), 397-401. doi:10.4172/1522-4821.1000397
- Arroyo, I., Woolf, B. P., Burelson, W., Muldner, K., Rai, D., & Tai, M. (2014). A Multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education*, 24(4), 387-426.
- Arsenio, W. F., & Loria, S. (2014). Coping with negative emotions: Connections with adolescents' academic performance and stress. *The Journal of Genetic Psychology*, 175(1), 76-90.
- Baker, R. S. J. D., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The Incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4), 223-241.
- Biswas, G., Segedy, J. R., & Bunchongchit, K. (2016). From design to implementation to practice a learning by teaching system: Betty's Brain. *International Journal of Artificial Intelligence in Education*, 26(1), 350-364.
- Burek, B. L. (2017). *Pilot study investigating the impacts of behavioural inattention and meta-attention on post-secondary students' online information seeking for academic purposes* (Unpublished master degree). University of Toronto, Canada. Retrieved from <http://hdl.handle.net/1807/79045>
- Chu, H. C., Tsai, W. W. J., Liao, M. J., Chen, Y. M., & Chen, J. Y. (2020). Supporting e-learning with emotion regulation for students with Autism Spectrum Disorder. *Educational Technology & Society*, 23(4), 124-146.
- Collins, R. (1990). Stratification, emotional energy, and transient emotions. In T. D. Kemper (Ed.), *Research agendas in the sociology of emotions*, (pp. 27-57). Albany, NY: State University of New York Press. Retrieved from <https://psycnet.apa.org/record/1990-97864-002>
- Conati, C. (2002). Probabilistic assessment of user's emotions in educational games. *Applied Artificial Intelligence*, 16(7-8), 555-575.
- Craig, S. D., Graesser, A. C., Sullins, J., & Gholson, B. (2004). Affect and learning: An Exploratory look into the role of affect in learning with an AutoTutor. *Journal of Educational Media*, 29(3), 241-250.
- Daradoumis, T., & Arguedas, M. (2020). Cultivating students' reflective learning in metacognitive activities through an affective pedagogical agent. *Educational Technology & Society*, 23(2), 19-31.
- D'Mello, S., & Graesser, A. (2013). AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(4), 1-39.
- D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2), 145-157.
- D'Mello, S., Taylor, R. S., & Graesser, A. (2007). Monitoring affective trajectories during complex learning. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (29) (pp. 203-208). Austin, TX: Cognitive Science Society. Retrieved from <https://escholarship.org/uc/item/6p18v65q>
- D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, 29, 153-170.

- D'Mello, S., Lehman, B., Sullins, J., Daigle, R., Combs, R., Vogt, K., Pekins, L., & Graesser, A. (2010). A Time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In V. Aleven, J. Kay & J. Mostow (Eds.), *Intelligent Tutoring Systems. ITS 2010. Lecture Notes in Computer Science* (vol. 6094, pp. 245-254). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-13388-6\_29
- Du, X., Zhang, M., Shelton, B. E., & Hung, J. L. (2019). Learning anytime, anywhere: A Spatiotemporal analysis for online learning. *Interactive Learning Environments*, (8), 1-15. doi:10.1080/10494820.2019.1633546
- Fida, R., Paciello, M., Tramontano, C., Fontaine, R. G., Barbaranelli, C., & Farnese, M. L. (2015). An Integrative approach to understanding counterproductive work behavior: The Roles of stressors, negative emotions, and moral disengagement. *Journal of Business Ethics*, 130(1), 131-144.
- Grafsgaard, J. F., Fulton, R. M., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2012). Multimodal analysis of the implicit affective channel in computer-mediated textual communication. In L. P. Morency, D. Bohus, H. Aghajan, A. Nijholt, J. Cassell, & J. Epps (Eds.), *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (pp. 145-152). New York, NY: Association for Computing Machinery. doi:10.1145/2388676.2388708
- Gross, J. J. (1998). The Emerging field of emotion regulation: An Integrative review. *Review of General Psychology*, 2(3), 271-299.
- Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, 85(2), 348-362.
- Haberman, S. J. (1979). *Analysis of qualitative data* (new developments, Vol 2, 1st ed.). Chicago, IL: Academic Press.
- Han, J., Zhao, W., Jiang, Q., Dong, Y., & Zhang, N. (2019). STEM intelligent learning environment and cognitive science research: An Interview with Prof. Gautam Biswas. *Open Education Research*, 25(2), 4-11.
- Harley, J. M., Bouchet, F., Hussain, M. S., Azevedo, R., & Calvo, R. (2015). A Multi-componential analysis of emotions during complex learning with an intelligent multi-agent system. *Computers in Human Behavior*, 48, 615-625.
- Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4), 470-497.
- Hwang, G. J., Sung, H. Y., Chang, S. C., & Huang, X. C. (2020). A Fuzzy expert system-based adaptive learning approach to improving students' learning performances by considering affective and cognitive factors. *Computers and Education: Artificial Intelligence*, 1, 100003. doi:10.1016/j.caeai.2020.100003
- Kobylińska, D., & Karwowska, D. (2015). How automatic activation of emotion regulation influences experiencing negative emotions. *Frontiers in Psychology*, 6, 1-4.
- Kołakowska, A., Szwoch, W., Szwoch, M. (2020). A Review of emotion recognition methods based on data acquired via smartphone sensors. *Sensors*, 20(21), 6367-6409. Retrieved from doi:10.3390/s20216367
- Lim, C. P. (2004). Engaging learners in online learning environments. *TechTrends*, 48(4), 16-23.
- Liu, T., & Lemeire, J. (2017). Efficient and effective learning of HMMs based on identification of hidden states. *Mathematical Problems in Engineering*, 1-26.
- Lu, J. (2012). The Research on the psychology of affective instruction. *Journal of Psychological Science*, 35(3), 522-529.
- Maheshwari, K., Nagendhiran, N. (2017). Facial recognition enabled smart doors using Microsoft face API. *International Journal of Engineering Trends and Applications (IJETA)*, 4(3), 1-4.
- Matsuda, N., Yarzebinski, E., Keiser, V., Raizada, R., Cohen, W. W., Stylianides, G. J., & Koedinger, K. R. (2013). Cognitive anatomy of tutor learning: Lessons learned with SimStudent. *Journal of Educational Psychology*, 105(4), 1152-1163.
- McDaniel, B., D'Mello, S., King, B. G., Chipman, P., Tapp, K. M., & Graesser, A. (2007). Facial features for affective state detection in learning environments. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 467-472). Austin, TX: Cognitive Science Society.
- 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(2014), 427-469.
- Patrut, B., & Spataru, R. P. (2016). Implementation of artificial emotions and moods in a pedagogical agent. In *Emotions, technology, design, and learning* (pp. 63-86). Academic Press. doi:10.1016/B978-0-12-801856-9.00004-9
- Pentel, A. (2015). Patterns of confusion: Using mouse logs to predict user's emotional state. In M. Kravcik, O. C. Santos, J. G. Boticario, M. Bielikova, T. Horvath (Eds.), *Proceedings of the 5th Preface International Workshop on Personalization Approaches in Learning Environments (PALE 2015)* (pp. 40-45). Dublin, Ireland: CEUR workshop proceedings. Retrieved from <http://ceur-ws.org/Vol-1388/PALE2015-complete.pdf>

- Popenici, S. A. D., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research & Practice in Technology Enhanced Learning*, 12(22), 1-13.
- Posner, J., Russell, J. A., & Peterson, B. S. (2005). The Circumplex model of affect: An Integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3), 715-734.
- Qin, J., Zheng, Q., & Li, H. (2014). A Study of learner-oriented negative emotion compensation in e-learning. *Journal of Educational Technology & Society*, 17(4), 420-431.
- Quinlan, J. R. (1990). Decision trees and decision-making. *IEEE Transactions on Systems, Man, and Cybernetics*, 20(2), 339-346.
- Raufelder, D., Regner, N., & Wood, M. A. (2018). Test anxiety and learned helplessness is moderated by student perceptions of teacher motivational support. *Educational Psychology*, 38(1), 54-74.
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145-172.
- Schutz, P. A., Pekrun, R., & Phye, G. D. (2007). *Emotion in education* (Vol. 10). doi:10.1016/B978-0-12-372545-5.X5000-X
- Segedy, J. R., Biswas, G., & Sulcer, B. (2014). A Model-based behavior analysis approach for open-ended environments. *Educational Technology & Society*, 17(1), 272-282.
- Taub, M., Mudrick, N. V., Azevedo, R., Millar, G. C., Rowe, J., & Lester, J. (2017). Using multi-channel data with multi-level modeling to assess in game performance during gameplay with Crystal Island. *Computers in Human Behavior*, 76, 641-655.
- Thompson, R. A. (1994). Emotion regulation: A Theme in search of definition. *Monographs of the Society for Research in Child Development*, 59(2-3), 25-52.
- Thornton, M. A., & Tamir, D. I. (2017). Mental models accurately predict emotion transitions. *Proceedings of the National Academy of Sciences*, 114(23), 5982-5987.
- VanLehn, K., Van De Sande, B., Shelby, R., & Gershman, S. (2010). The Andes physics tutoring system: An Experiment in freedom. In R. Nkambou, J. Bourdeau, & R. Mizoguchi (Eds.), *Advances in Intelligent Tutoring Systems* (pp. 421-443). Berlin Heidelberg: Springer.
- Wang, S., Wu, H., Kim, J. H., & Andersen, E. (2019). Adaptive learning material recommendation in online language education. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Proceedings of 20<sup>th</sup> International Conference on Artificial Intelligence in Education* (pp. 298-302). Cham, Switzerland: Springer.
- Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-aware tutors: Recognizing and responding to student affect. *International Journal of Learning Technology*, 4(3-4), 129-164.
- Yu, L. C., Wu, J. L., Chang, P. C., & Chu, H. S. (2013). Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowledge-Based Systems*, 41, 89-97.
- Zhang, T. (2008). *Research on learning emotions and emotion models under the network teaching environment* (Unpublished doctor dissertation). Capital Normal University, China. Retrieved from <https://kns.cnki.net/kcms/detail/detail.aspx?FileName=2008144384.nh&DbName=CMFD2009>

## Exploring the Relationships between Achievement Goals, Community Identification and Online Collaborative Reflection: A Deep Learning and Bayesian Approach

Changqin Huang<sup>1,2\*</sup>, Xuemei Wu<sup>2</sup>, Xizhe Wang<sup>1</sup>, Tao He<sup>2</sup>, Fan Jiang<sup>1</sup> and Jianhui Yu<sup>1</sup>

<sup>1</sup>Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Zhejiang Normal University, Jinhua, China // <sup>2</sup>School of Information Technology in Education, South China Normal University, Guangzhou, China // cqhuang@zju.edu.cn // wuxuemei@m.scnu.edu.cn // xzwang@zjnu.edu.cn // tao.he2016@gmail.com // jiangfan1116@foxmail.com // jianhuiyu@126.com

\*Corresponding author

**ABSTRACT:** Collaborative reflection (co-reflection) plays a vital role in collaborative knowledge construction and behavior shared regulation. Although the mixed effect of online co-reflection was reported in the literature, few studies have comprehensively examined both individual and group factors and their relationships that affect the co-reflection level. Therefore, this study explored the structural relationships between achievement goals (task-based, self-based, and other-based goals), online community identification, and co-reflection, which can consequently assist instructors in improving the related pedagogical strategies. To this end, 26813 posts on MOOC and college online learning platforms were gathered. Specifically, deep learning techniques were first used to train a classifier that classifies the large-scale co-reflection text automatically. The Bayesian method was then applied to disclose the structural relationships among achievement goals, community identification, and co-reflection. The results showed that the proposed classification algorithm achieved the best performance. Two best-fit models for characterizing the respective relationships between co-reflection and community identification as well as achievement goals were obtained using the Bayesian method. The results of the experiments on these two models demonstrated that both task-avoidance and other-avoidance goals were related directly to co-reflection, all task-approach, self-approach and other-approach goals were related indirectly to co-reflection, but self-avoidance goals had both a direct and an indirect relationship with co-reflection. The relationship between community identification and co-reflection was mediated by other-based goals. Some theoretical and practical implications were discussed for instructors and practitioners to build an online community.

**Keywords:** Deep learning, Bayesian network, Achievement goals, Co-reflection, Community identification

### 1. Introduction

Co-reflection refers to a process of collaborative critical thinking and knowledge construction, the activities of which are commonly affected by a combination of elements of individuals and groups (Kalk, Luik, & Taimalu, 2019). One way of supporting co-reflection is to use the tools provided by information and communication technology, such as blogs, e-portfolios, Facebook etc. In particular, these tools accommodate an open, flexible and diverse online learning community where students can reflect collaboratively on their thoughts, compared to expressing their thoughts in traditional ways (Yilmaz & Keser, 2016). Individuals would be motivated to read other peers' postings and comments, to develop a sense of community. In turn, more time is spent on their postings, which consequently may lead to an in-depth reflection (Clarà, Kelly, Mauri, & Danaher, 2017; Huang, Han, Li, Jong, & Tsai, 2019). However, researchers have assessed the level of online co-reflection with reporting mixed results. Some studies have shown that many students only describe or summarize what happened rather than critically think about it (Ozkan, 2019). Dalgarno, Reupert, and Bishop (2015) stated that some negative responses are given due to the lack of peer feedback, apparent resistance, and learning community engagement etc. However, few studies have investigated the antecedents and driving mechanism of online co-reflection, which can provide some theoretical and practical implications to motivate learners to be deeply engaged.

Previous researchers explored factors influencing co-reflection such as peer feedback and interactive behavior (Novakovich, 2016). However, individuals' participation in communities is for certain purposes, however learning motivation refers to some significant individual factors that guide and regulate individuals' behavior (Lim & Lim, 2020), which is the condition of intention to act (Chang, Hou, Wang, Cui, & Zhang, 2020). Therefore, there is a strong need to further investigate the factors influencing co-reflection from the perspective of motivation. Community identification is another crucial concept that facilitates members' participating, sharing, and knowledge constructing (Ergün & Avcı, 2018). It also plays a significant role in bridging the individual and group factors (Chang et al., 2020). Some studies have indicated that there may be different interactive relationships between community identification and achievement goals in a collaborative environment

(Chang et al., 2020; Thijs & Fleischmann, 2015). Therefore, this study was designed to explore the relationships among different achievement goals, community identification, and co-reflection in an online learning community.

In addition, the large-scale online discussion data and reflective writing provide valuable information to understand students' co-reflection, but also raise some problems of data analysis (Liu, Zhang, Wang, & Chen, 2017). Although features can be automatically captured from the data by machine learning methods, costly manual engineering is also required (Ullmann, 2019). Deep learning is a representation learning technique which can process the raw input to be suitable for the classification of feature engineering, and it has been recognized as the most advanced solution to performing tasks in data mining related to classification (LeCun, Bengio, & Hinton, 2015). However, few works have applied deep learning techniques to analyze reflective texts (Chen, Xie, Zou, & Hwang, 2020).

For this research, the deep learning technique and Bayesian method are applied to make the automatic prediction of online co-reflection levels, as well as discover the relationships between achievement goals, community identification and co-reflection. Specifically, two research questions (RQ) are proposed in this study:

RQ1: To what extent can the deep learning technique accurately classify the level of co-reflection of each student?

RQ2: What are the relationships between achievement goals, community identification and co-reflection?

## **2. Theoretical background**

### **2.1. Achievement goal theory**

Achievement goal theory is a predominant theoretical framework of achievement motivation to interpret different qualities of individual learning and well-being, particularly in educational contexts (Urda & Kaplan, 2020). Various models of the achievement goal theory have been proposed to conceptualize students' motivational orientations to understand students' motivational beliefs, their causes and effects (Elliot, Murayama, & Pekrun, 2011; Elliott & Dweck, 1988). Existing studies emphasized that learning motivation and achievement goals provided an essential foundation for reflection and meaning construction (Anderman, 2010; Tikhomirova & Kochetkov, 2018). Some researchers indicated that learners might have diverse goal-oriented motivation mechanisms in different contexts, e.g., individual versus collaborative learning environments (Lim & Lim, 2020). Thijs and Fleischmann (2015) pointed out that achievement goals depended on individuals' perception of relatedness to others. Therefore, this research will further explore the driving mechanism of achievement goals on co-reflection in a collaborative learning community.

### **2.2. Social identity theory**

Social identity theory (SIT) provides an essential theoretical background for community identification and member behavior, which indicates that group members establish their identity in a community by viewing themselves as a part of that, and generating an emotional attachment to the group or community (Tajfel, 1978). It should be noted that social identification involves not only perceived self-categorization, but also the evaluative and affective states with the social group, and this identification with the group allows members to modify their thoughts and behaviors (Qu & Lee, 2011). Chang et al. (2020) found that community identification significantly mediated the relationship between motivation and members' community behavior. Additionally, Bowskill (2017) pointed out that inducing a sense of group identity can motivate self-evaluation and critical thinking engagement within a technology-supported learning community. Therefore, the learners' sense of identity with the group might be an important factor influencing co-reflection in this study.

Further, considering that co-reflection is a process of knowledge co-construction, including individual and group cognition, it is necessary to comprehensively investigate the essential individual and group factors that affect it. Grounded on achievement goal theory and community identification theory and the related research, this study mainly focuses on two pivotal factors, achievement goals and the learners' sense of identity with the group and reveals their driving mechanism for online co-reflection.



### **3. Literature review**

#### **3.1. Co-reflection**

Co-reflection is a process of collaborative critical thinking involving cognitive and affective interactions between two or more individuals who explore their experiences to reach new intersubjective understandings and appreciations (Yukawa, 2006). This definition of co-reflection brings new perspectives and considerations from the dialogue with others who might see situations differently, challenge assumptions, or ask significant questions (Krutka, Bergman, Flores, Mason, & Jack, 2014). These arguments are consistent with those by Vygotsky (1978) who assumed that cognition is a process of social interaction with each other. In this study, we also believe that co-reflection would be deepened when engaged in communion with peers who could push each other beyond description to thoughtful reconsideration (Krutka et al., 2014). However, existing studies mainly explored platforms or strategies that support co-reflection. Kalk, Luik, and Taimalu (2019) reported that the reflection level can be predicted by the characteristics of students, blog groups and blogging. But the essential factors that affect the level of co-reflection and its driving mechanism are still lacking.

#### **3.2. Achievement goals**

Achievement goals are the integrated systems, theories, or schemas, that incorporate conceptions of ability, perceptions of the self and features of self-consciousness, definitions of success in specific achievement contexts, and affective and behavioural responses (Urda & Kaplan, 2020). Recently, the latest achievement goal theory model proposed by Elliot, Murayama, and Pekrun (2011) offers a six-component model, which includes task-approach, task-avoidance, self-approach, self-avoidance, other-approach, and other-avoidance. And all of them are distinguished by task, self, and other three competence evaluation standards. Elliot and Thrash (2001) remarked that six possible types of achievement goals as the basis for evaluation have many benefits, that is, it explicitly accounts for both the energization and direction of competence-based behaviour, and provides a more specific definition of the achievement goal construct. Also, it affords greater conceptual flexibility in that any combination of reason and goal may be considered. Therefore, this model was adopted as one conceptual framework in the present study.

Additionally, although several studies have been conducted to explore the relationship between achievement goals and reflection (Mercier, 2017), a consensus was not reached about the effects of different achievement goals on reflection (Urda & Kaplan, 2020). Moreover, studies on the relationship between achievement goals and reflection mainly focus on individual reflection (Collin & Karsenti, 2011). Thus, it is meaningful to conceptualize the effects and driving mechanism of different achievement goal orientations on co-reflection in an online learning community.

#### **3.3. Community identification**

According to social identity theory, community identification refers to the degree to which individuals feel a sense of belonging to the community (Tajfel, 1978). Feeling like part of the group in a community is considered a critical factor for a successful online community building (Qu & Lee, 2011), and members with a high level of identification can reduce their stress, enhance their self-esteem and be motivated to modify their thoughts and behaviors according to the group's common values and interests (Chiu, Huang, Cheng, & Sun, 2015). Recently, attention was given to this potential pathway that links community identification and community participation, members' knowledge sharing and construction (Yilmaz, 2016). Thus, exploring the relationship between different achievement goals and community performance will help us better understand the individual's behavior.

#### **3.4. The relationships between achievement goals, community identification and co-reflection**

Previous studies have explored the relationship between achievement goals and co-reflection, showing that there are different direct and indirect relationships between them. Mercier (2017) found that although learning and performance goals displayed no differences in outcome measures, groups with the former goal showed more reflection and explanations than groups with the later goal during the task. Lau, Liem, and Nie (2008) reported that task-approach and task-avoidance goals have both a direct and an indirect effect on deep learning, and the relationship between the two of them and deep learning is mediated by classroom attentiveness and group

participation. However, group participation mediated the relationship between the performance-approach goal and deep learning. Chang et al. (2020) pointed out community identification significantly mediated the relationship between motivation and social loafing. Therefore, it can be inferred that the relationship between achievement goals and co-reflection may be mediated by community identification, and different goal orientations may have different indirect or direct relations to co-reflection.

Conversely, some studies indicated that the other different potential path exists between achievement goal, community identification and co-reflection. For example, Zumbunn, McKim, Buhs, and Hawley (2014) found that expectancy (one of the important motivational constructs) significantly mediated the relationship between sense of belonging (a construct like community identification) and achievement. But task values failed to mediate the relations. Further, Won, Wolters, and Mueller (2018) examined the relationships between sense of belonging, achievement goals and self-regulated learning, reporting that only mastery goals mediated the relationship between the sense of belonging and metacognitive. This implies that students' identification affects the achievement goals or the reasons or purposes they used in the task, which in turn impact their academic effort and engagement (Won et al., 2018). The self-determination theory (SDT) can provide some supportive evidence for this, which underlined the need for relatedness to others plays a critical role in students' motivation and performance (Van den Broeck, Ferris, Chang, & Rosen, 2016). Therefore, different goal orientations will be affected by community identification with varying degrees.

Taken together, there may be two different potential relationships between achievement goals, community identification and co-reflection. However, further two important gaps need to be noted and filled. First, although existing research investigated the relationship between achievement goals and reflection, the accurate relationships between different goal orientations, community identification and co-reflection are still unknown. Therefore, this study explored the relationship between co-reflection and achievement goals based on the six-factor achievement goals model. Second, prior studies mostly proposed a hypothetical model and used the structural equation modelling method to further verify the fitting effect, which is theory-driven. Instead, this study attempts to mine the relationships between different achievement goals, community identification and co-reflection using the Bayesian method from a data-driven perspective.

### **3.5. Deep learning for educational applications**

Deep learning has a multilayer network structure and has a strong power to learn discriminative information from examples, patterns or events (Waheed et al., 2020). Many applications, such as learning performance prediction, learning recommendation, intelligent learning tool and system development, have been explored based on various methods (Hwang, Sung, Chang, & Huang, 2020; Hwang, Xie, Wah, & Gašević, 2020; Wang, Mei, Huang, Han, & Huang, 2021; Zhou, Huang, Hu, Zhu, & Tang, 2018). The most commonly used method is text classification in educational data mining (Chen, Xie, & Hwang, 2020; LeCun et al., 2015). Ullmann (2019) concluded that there are three approaches (machine learning-based, dictionary-based and rule-based) for reflective text analysis. However, all of these have their limitations (e.g., costly manual feature engineering, time-consuming etc.). Deep learning has great potential for educational data mining, especially in text classification (Young, Hazarika, Poria, & Cambria, 2018). Therefore, deep learning is conducted for co-reflection text classification in this study.

## **4. Methodology**

### **4.1. Research design**

To answer the two research questions, this study consists of four stages, as depicted in Figure 1. Specifically, students' online co-reflection text data and the questionnaire data of achievement goals and community identification was collected. Furthermore, the text and questionnaire data were further preprocessed to ensure validity. For RQ1, this study adopted the techniques of BERT and LSTM to classify reflective texts to identify students' co-reflection level, then the performance of the classification model was evaluated. For RQ2, the trained classification model was used to identify each student's level of co-reflection, and the Bayesian method was then integrated to explore the relationship between the three factors (online co-reflection, achievement goal, community identification).

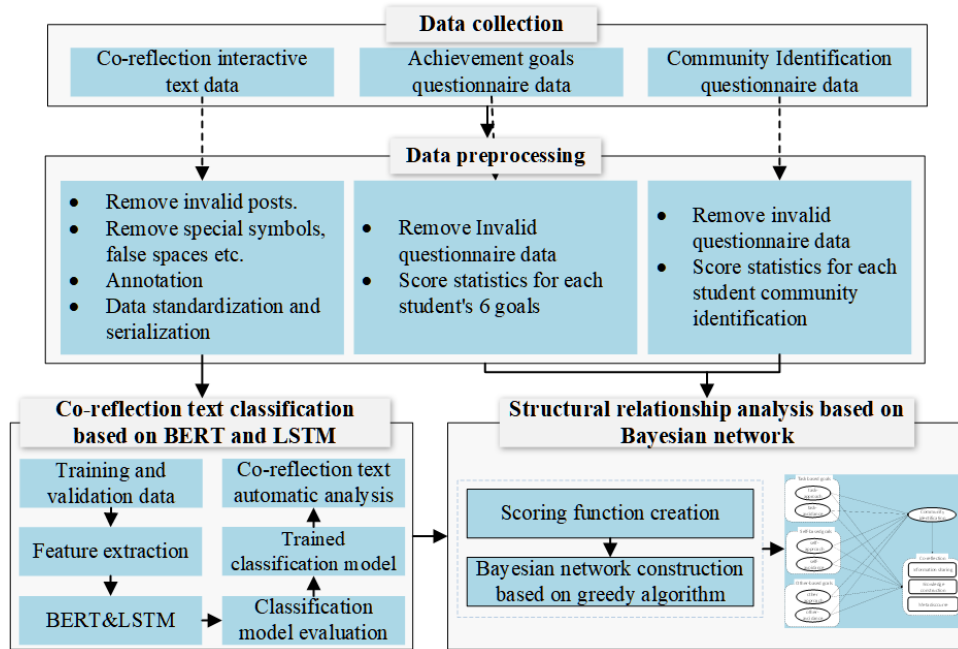


Figure 1. The research design of this study

#### 4.2. Data collection and preprocess

The co-reflection text data was collected from the three online courses on educational technology, with each course being offered for almost five months. During the course, learners participated in online co-reflection activities in a similar way that each discussion and reflection began after the topic was posted. From a total of 26813 original posts collected, 16890 posts were determined as the dataset after removing the invalid data. To reduce noise in the dataset, all duplicate posts and special symbols such as punctuation marks, false spaces and emoticons were removed according to Liu's et al. (2017) recommendation.

The data for the other two variables under consideration were gathered through a Chinese survey website. In the survey, a total of 115 undergraduate and graduate students who participated in an average of 12 to 19 online co-reflection activities in the two courses were invited to complete the questionnaires voluntarily. In the first class of the course, students were invited to fill in the achievement goal questionnaires adopted from Elliot et al. (2011) which comprised 18 measurement items. At the end of the course, students filled in the community identification questionnaires adopted from Chang et al. (2020) comprising four measurement items. All the items were measured on a five-point Likert scale ranging from "not true of me (1)" to "extremely true of me (5)." Finally, a total of 95 valid responses were obtained. The Cronbach's alpha of achievement goal and community identification were 0.916 and 0.898 respectively.

To train an efficient classifier, a structural dataset with label information was constructed for training and testing the classification model. To do this, the unit of analysis was defined as a complete dialogue with the same peer on each topic, also called an episode (Mercier, 2017). Each unit of the analysis was coded by two coders according to Lei and Chan's (2018) coding scheme. Specifically, the scheme consists of nine reflection levels (see Table 1), in which 1 to 3, 4 to 6, and 7 to 9 are reconsidered as low, middle, and high three levels of sharing of information, knowledge construction, and metadiscourse, respectively. In this study, each analysis unit was marked as one of three levels 1, 2, and 3 respectively, with a unit that does not belong to any of the nine categories marked as 0. Discussions and revisions were undertaken among the research team members until consensus was reached on each post. Finally, Cohen's Kappa was computed as 0.878 ( $p < .01$ ), which indicates a high level of agreement between the coders.

Table 1. The coding scheme of co-reflection levels

Categories	Description	Exemplar excerpts from co-reflective logs
1. Listing and Copying	Lists notes without explanations; copies information from or repeat other's notes in a very close way	Share an article "coupled teacher" or "double loss," see the link below.
2. Brief Summary	Summarizes a few notes shortly and often incompletely	By summarizing the views in the two articles, the principles of CAI courseware design are as follows: 1. Educational principles 2. The principle of control...
3. Interpretation or Elaboration	Interprets others' notes on information with different wording or extends information by examples or evidence	The previous students mentioned many professional tools, and almost gave a detailed overview of..., I still tend to recommend the two most commonly used tools, blog and WeChat...
4. Question-Based Discussion	Sees the discussion as question-based and a deepening process of seeking answers to questions	In my view, the focus of educational technology is technology, I think educational technology is ...
5. Constructive Use of Information	Uses information, either from experts, books, the Internet, or other related courses, life experience, etc. to justify or deepen ideas	Once we visited the teacher, he suggested that we should pack the knowledge of each chapter and put the packed knowledge in different boxes...I think the process of finding and marking boxes is the process of building knowledge scaffolding, because...
6. Intertwined Question Explanation	Keeps asking related questions, showing doubt or seeking clarification; responses and explanations are intertwined progressively in the discussion	Can the cultivation of innovation ability be reflected through their group discussion process? ... For example, encourage them to innovate in the display of the discussion results, etc.
7. Meta-Cognition	Reflects on what the class does not know; realizes high points in the discussion; self-defines goals and tasks for exploration	Our current progress is to learn some artificial intelligence knowledge, I think the purpose is to be able to understand the relevant papers. The first step of the next plan is to improve academic literacy
8. Meta-Theory	Focuses on theories while developing the discourse; uses theories/conjectures to explain the phenomena, even making attempts to create new theories	...When I mentioned how to balance curriculum planning, I thought of Cuba's thought-provoking point...It must be explained that the emergence of information technology has raised the issue of curriculum design... Therefore, education and technology themselves are also a pair of balanced propositions.
9. Meta-Conversation	Focuses on examining what the discourse is about, especially reflecting on discourse goals; adopts a "we" perspective to assume collective responsibility for advancing knowledge; tackles difficult/important issues which may be neglected by the community	Yes, there is a discussion that can produce a collision of ideas...So the purpose of the mutual evaluation is designed to urge the group members to participate in the group discussion more seriously.
10. Other	Some posts include greetings, thanks, simple compliments, etc.	Very good! Thank you! Morning! etc.

### 4.3. Co-reflection text classification based on BERT and LSTM

Previous studies have shown that it is difficult for students to achieve a deep level of reflection in a short time, and the quality of the reflection is related to the mastery of knowledge (Granberg, 2010; Van den Kieboom, 2013). Differing from the existing classification models, the long short-term memory (LSTM) model that can capture long-term dependencies (Yu, Si, Hu, & Zhang, 2019) was therefore employed to obtain the time series information of the reflective text. The BERT (Bidirectional Encoder Representations from Transformers) pre-

training model performs the best on language understanding and text extraction (Devlin, Chang, Lee, & Toutanova, 2018). Therefore, BERT and LSTM were integrated to classify co-reflection levels from a large-scale dataset. The overall BERT and LSTM architecture of our classification model can be seen in Figure 2. A total of 10572 labelled posts were used as the training dataset and the reflective text data of each student was arranged in chronological order of reflection topics (Topic1, Topic2, ...Topic  $m$ ). Each post was segmented and vectorized based on the jieba library as well as BERT's pre-training. Vectorized and positioned co-reflection text information was obtained and used as input to the BERT model for fine-tuning. In this way, a serialized vector of the co-reflection text from each topic was obtained ( $C_1, C_2, \dots, C_m$ ). It was later taken as the input of the text classifier based on the LSTM model. Finally, a fully connected (FC) layer and Softmax function were used to classify the output vector of LSTM and generate the final prediction result.

For training, ten-fold cross-validation was used for each algorithm as well as the metrics of accuracy, precision, recall, and F1 (the harmonic mean of precision and recall) which are commonly used to evaluate the performance of text classification tasks and measure the proportion of correct predictions from different perspectives (Hew, Hu, Qiao, & Tang, 2020). Therefore, these metrics were employed to measure the performance of the classification model in this study. Other pre-training models (e.g., Word2vec), serialization analysis methods (e.g., historical average (HA)), and keywords methods (e.g., TF-IDF) were also implemented for the classification task in the present research. After the training process, the text classification model was used to automatically classify the rest of the texts into different levels of co-reflection. The resulting levels will be used for structural relationship analysis presented in the next section.

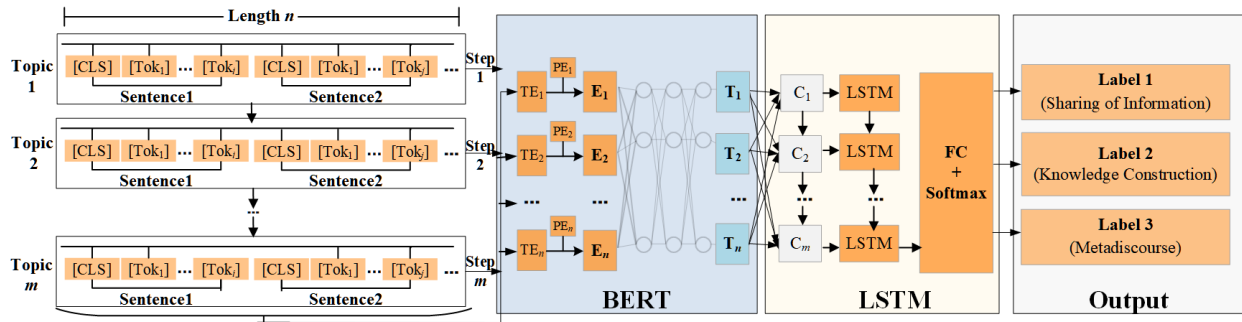


Figure 2. Co-reflection level prediction model integrating BERT and LSTM

Note.  $m$  = the total topic length;  $n$  = the number of words in each topic; TE = text embedding, and PE = the position embedding of text.

#### 4.4. Structural relationship analysis based on Bayesian networks

The Bayesian method of directed association mining was used to explore the relationships between achievement goals, community identification and co-reflection. Traditional confirmatory data analytic procedures (e.g., path analysis, structural equation modelling) follow a “frequentist” approach to test a network of effects in a model. Such approaches generally do not fit well with the prescribed model, which may lead to improving the model fit by practice with inherent problems (Hagger & Hamilton, 2018). Instead, the Bayesian approach assumes that model parameters have inherent uncertainty that is represented by a distribution. As a powerful tool that infers uncertain association relationships, the Bayesian approach has been widely used to automatically mine the associations and causal relationships between factors (Heckerman, 1997). It is well suited for exploring the relationships between achievement goal orientations, community identification and co-reflection in this study. However, some researchers recommend that caution should be taken and more prior information ought to be considered when using the Bayesian approach (Hagger & Hamilton, 2018; Meyer & Xu, 2007). Therefore, in this study, the theoretical prior knowledge is comprehensively considered with the Bayesian structure analysis.

To determine the optimal structural relationship between the three factors, eight nodes in a Bayesian network were constructed firstly, and then the most appropriate network structure was selected from the existing datasets. The distinctions between six orientations of goals have been validated by multiple studies (Elliot et al., 2011). In this study, the internal logical relationships between the six types of goals were therefore regarded as controlled. Achievement goals, community identification, and co-reflection were then set to the first, second, and third levels of the network, respectively. After that, an optimal model would be determined based on the existing data. Then, the positions of achievement goal and community identification in the network were swapped, and the same operation was repeated for others.

To improve the model further, the scoring function method was implemented to evaluate the degree of fitting between the Bayesian network and training dataset. As such, whether to add, remove or adjust the directions of the edges of the Bayesian network was determined by looking at the changes of the score. Note that the nodes represent the variables and the edges indicate the relationship between two variables in the Bayesian network. That is, changes to the edges are equivalent to exploring possible relationships between the variable. Specifically, in Figures 3 and 4, the BDeu (“BD” for Bayesian Dirichlet, “e” for likelihood-equivalence, “u” for uniform joint distribution) score, K2 score and Bic score measure the degree of fit. The larger the value, the better the model fits (Carvalho, 2009). In addition, a greedy algorithm was used to identify a stable relationship structure by continually updating until the score function value remains unchanged. In this, a relatively stable network relationship structure can finally be obtained.

## 5. Results

### 5.1. Co-reflection text classification results for RQ1

Table 2 lists the performance results of the classification models with different algorithms. The classification model which integrates BERT and LSTM performed better than the other models. Specifically, the pre-training model based on BERT (e.g., BERT & LSTM, BERT & HA) performed better than Word2Vec (e.g., Word2Vec & LSTM, Word2Vec & HA), and TF-IDF performed the worst. Furthermore, the algorithms that integrated the pre-training model and the serialization model (e.g., BERT & LSTM, Word2Vec & LSTM) performed better than those without the serialization model (e.g., BERT & HA, Word2Vec & HA). Again, the algorithms that combine BERT and LSTM performed the best. However, the results also revealed that these algorithms were less effective than human judgments. In particular, the performance of our algorithm in terms of precision, recall, accuracy, and F1 was an average of 3.7% lower than human judgments. According to the literature, the error was acceptable within 10% (Ullmann, 2019). Generally, the classification model of integrating BERT and LSTM demonstrated reasonably good performance.

Table 2. Text classification model results

	Precision	Recall	Accuracy	F1
Human	81.25%	78.00%	78.95%	79.63%
TF-IDF	58.33%	58.33%	57.89%	58.33%
Word2Vec & HA	62.50%	63.83%	63.16%	63.16%
Word2Vec & LSTM	66.67%	69.57%	68.42%	68.12%
BERT & HA	64.58%	70.45%	68.42%	67.52%
BERT & LSTM	75.00%	76.60%	75.79%	75.80%

### 5.2. Structural relationship results for RQ2

Two models were chosen (see Figures 5 and 6) through multiple rounds of evaluation and selection. Figure 3 and Figure 4 show the trend of the BDeu, K2, and Bic scores of the two models respectively as the number of edges of the model decreased. As shown in Figure 3, as the number of edges in the model decreases, the score of BDeu becomes larger. But BDeu and Bic scores tend to be flat when five edges in this model have been removed. Continuously, an obvious downward trend of the K2 score was observed when six edges in the model have been removed. This indicates the fit degree between the model and data is relatively higher without the need to provide more information. Taken together, an approximate optimal model (model 1) was obtained. In the same way, model 2 was also obtained.

For model 1, all eight variables were selected from the competing admissible models. The conditional probability of each variable was computed according to the standardized values (0, 1, 2) converted from the original scores of achievement goals and community identification variables. It measures the degree of the links between different variables. According to the selected model 1, this shows that community identification mediated the relationship between achievement goals and co-reflection. Specifically, both the task-avoidance goal and other-avoidance goal have direct relations to co-reflection, while the three goals of the task-approach, self-approach and other-approach have an indirect link to co-reflection. Interestingly, the self-avoidance goal has both a direct and an indirect path to co-reflection.

For model 2, eight variables were similarly retained after the selection. Specifically, this admissible model showed that community identification has only indirect connections to co-reflection, and the relation is mediated by other-based goals. The task-avoidance goal and self-avoidance goal only has a direct relation to co-reflection, respectively, whereas the goals of task-approach and self-approach have no direct relation to co-reflection.

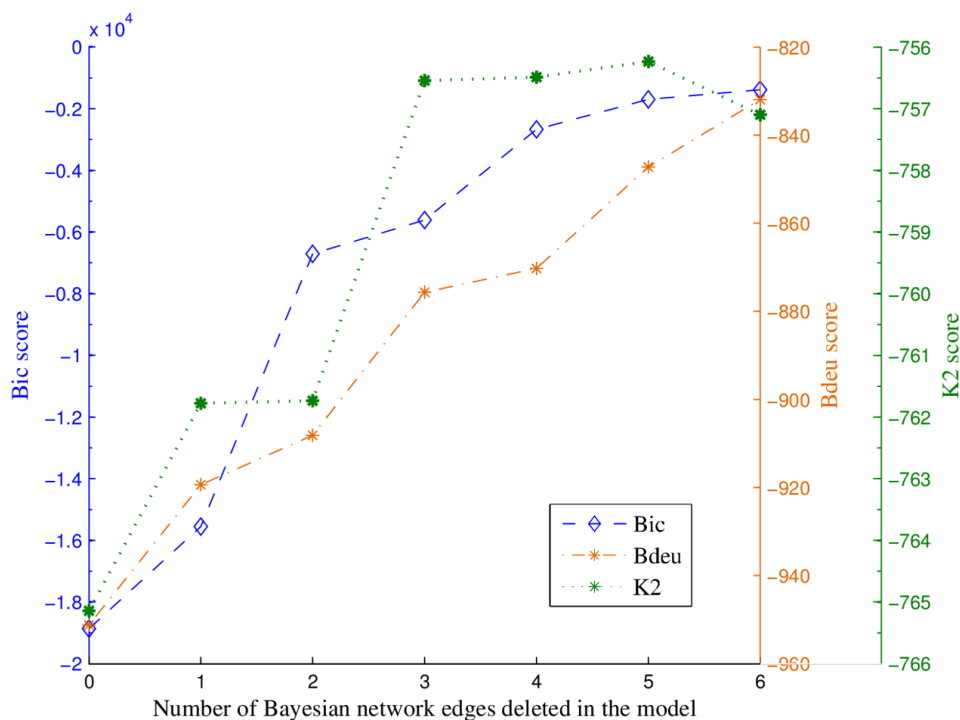


Figure 3. Bdeu, K2, Bic score in model 1

Note. 0-6 in the figure in the x-axis is the number of Bayesian network edges deleted in the model. The y-axis is the score.

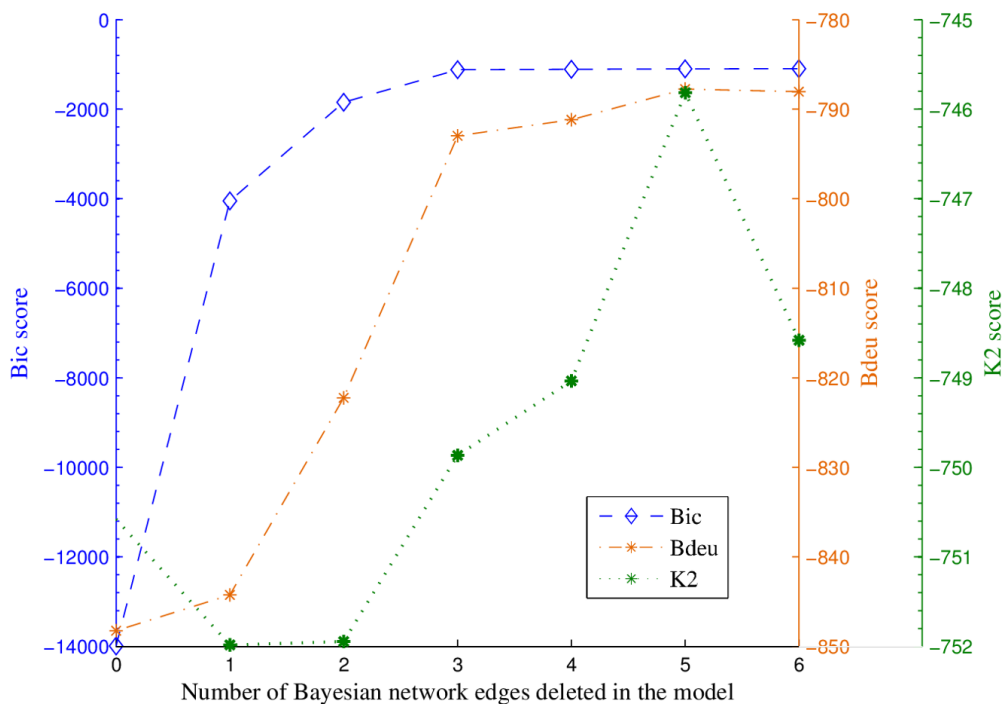


Figure 4. Bdeu, K2, Bic score in model 2

Note. 0-6 in the figure in the x-axis is the number of Bayesian network edges deleted in the model. The y-axis is the score.

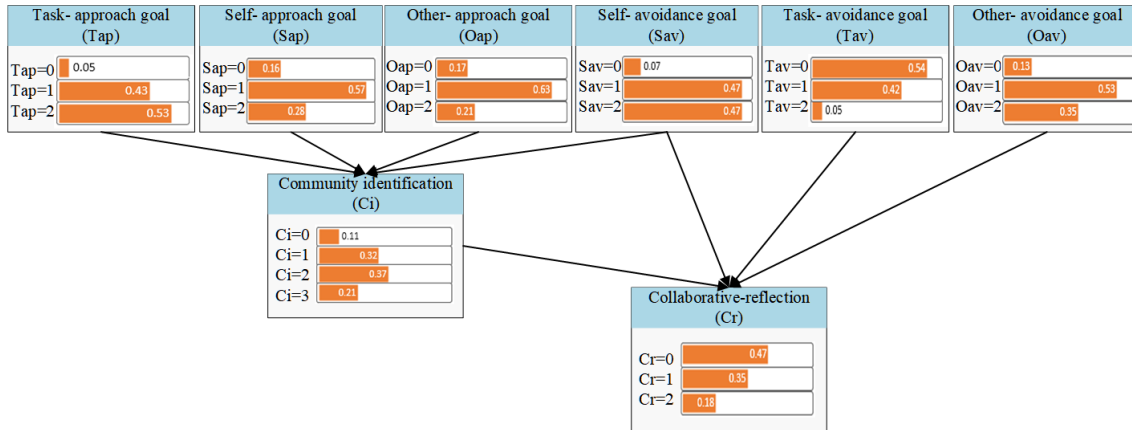


Figure 5. The structural relationship of model 1

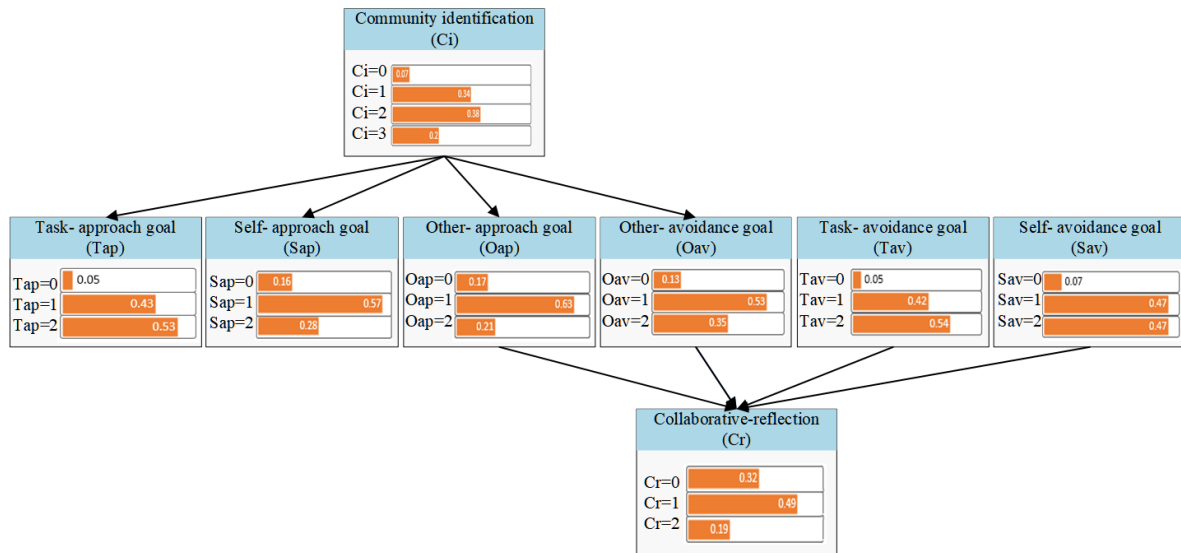


Figure 6. The structural relationship of model 2

## 6. Discussion

To analyze the online large-scale interactive text data about students' co-reflection information, the present study combined the BERT pre-training model and LSTM into an integrated classifier that performed better than the baseline models. On the one hand, the BERT pre-training model uses the mask method and has migration capabilities (Devlin et al., 2018) which can quickly and precisely understand the feature of reflective text language in this study. On the other hand, the BERT model with the embedded attention mechanism, which is not limited to the length of the text sequence, can improve the accuracy of the classification model compared to conventional methods (González-Carvajal & Garrido-Merchán, 2020). In addition, the time series feature of the reflective text is captured based on LSTM, which is in line with the actual development of the learner's level of reflection and accords with reflection as the essential feature of the internal cognitive process (Granberg, 2010). Therefore, the integrated classification model can more accurately identify the co-reflection level and this model also confirmed the advantages of using deep learning techniques for educational data mining, especially for text classification tasks (Young et al., 2018).

Along with the two best-fit models found in this study, the different potential relationships between achievement goals, community identification and co-reflection were indicated. For task-based goals, task-approach goals were not directly related to co-reflection, and community identification mediated the links of the task-approach goal and co-reflection. This is not completely consistent with the existing conclusion that mastery goals were both directly and indirectly related to deep learning strategies and outcomes (Heo, Anwar, & Menekse, 2018; Lim & Lim, 2020). There are a few possible explanations for this inconsistency. First, the students were required to participate in co-reflection activities that interacted with others, and this may push them to reach intersubjective understandings. That is, in this process, they would have sense of community, which in turn



affected their co-reflection further. This conforms to previous studies (Lau et al., 2008). Another possible explanation is that the community identification perceived by the students may increase their task value (David, 2014), which in turn promoted their participation in co-reflection. This assumption was also made by Zumbrunn et al. (2014), but further investigation is still needed. Furthermore, for task-avoidance goals, positive or negative relationships between task-avoidance goals and help-seeking behaviors have been described in previous studies. Based on the present results, relatedness to others, however, may not be the main psychological need that motivated these students to work hard and participated in co-reflection (Van den Broeck et al., 2016). According to Elliot et al. (2011), students with task-avoidance goals mainly attained satisfaction by completing challenging tasks. Therefore, further investigations on considering the factor of the task value may be more helpful to understand the relationships between task-based goals, community identification, and co-reflection.

For other-based goals, the result indicated that the depth of co-reflection for the students was mainly affected by their perceived community identification. It may regulate their learning strategies and goals for participating in co-reflection. This is not entirely consistent with the existing conclusion (Lau et al., 2008; Won et al., 2018). But as for the psychological need and competence evaluation criteria of the task-approach goal, the results implied in the present study is consistent with existing findings (Elliot et al., 2011; Van den Broeck et al., 2016). Besides, community identification did not mediate the relationship between the other-avoidance goal and co-reflection. According to Payne, Youngcourt, and Beaubien (2007), students with other-avoidance goals had low help-seeking behavior and a sense of efficacy. They may be afraid of showing incompetence in front of their peers, with an attitude of resistance and avoidance to the community. Therefore, community identification would not mediate its relations to co-reflection unless the community identification was enough to allow them to regulate their own goals, and the performance of co-reflection could be promoted. Overall, students with other-based goals would regulate their goals through community identification in a way that affected their performance on co-reflection. According to SDT, different from the task-based goal, students with other-based goals may mainly take relatedness to others as their main psychological needs (Van den Broeck et al., 2016). But they may have different ways of behavioral regulation. For students with the other-avoidance goal, autonomous regulation and controlled regulation were dominant, while students with the other-approach goal possibly had controlled regulation (Deci & Ryan, 2012).

For self-based goals, students with self-approach goals had higher internal motivation and help-seeking behaviors (Elliot et al., 2011). Therefore, if they received help from their group, they may have a higher level of a sense of belonging. This could encourage them to share their knowledge and promote co-reflection. This fact was in line with the principle of cooperative reciprocity and the claims of SIT (Chiu et al., 2015). Unlike the self-avoidance goal, students with the self-approach goal, however, can regulate their own goals due to their perceived community identification, which was inconsistent with the finding of Elliot (Elliot et al., 2011). We inferred that the self-approach goal followed the competence evaluation criteria of self-improvement, but one would have more satisfaction and efficacy, tending to be in line with the group's common values and interests after an individual's goals were accepted by their group (Chiu et al., 2015). Moreover, the self-avoidance goal had a direct and an indirect link to co-reflection, but it was not affected by community identification. This conformed with Elliot's et al. (2011) view. The self-avoidance goal was based on self-improvement as the competence evaluation standard, and they feared performing worse than they had performed before. Therefore, their goals would not change due to the community identification they felt.

## 7. Conclusion

This research has comprehensively examined the factors that affect online co-reflection. In particular, the different relationships between achievement goals, community identification, and co-reflection were revealed using deep learning techniques and Bayesian methods. This work has made the following contributions. First, the present study is one of the few works that applies deep learning techniques to classify reflective texts to identify the learner's co-reflection level, which provides a methodological foundation for the construction of a platform which automatically monitors learners' co-reflection level. Second, this study has further validated the six-factor achievement goal framework by demonstrating the significance of the achievement goal theory in the context of online collaborative learning. Third, some practical implications can be provided for online community builders and instructors according to the driving mechanism of co-reflection found in this research. Specifically, to promote learners' in-depth co-reflection, practitioners should comprehensively consider learners' achievement goal orientations and community identification for providing the corresponding guidance.

The study also has several limitations. First, the data were collected using different methods. This may lead to deviations among different evaluation standards, affecting the accuracy of the results to some degree. Second,

this research does not consider the causal relationship between different factors, using an exploratory attempt instead. Third, this research mainly focuses on theoretical and methodological exploration, but lacks practical educational applications.

There are some possible directions for extending this research. First, multi-modal and longitudinally serialization data should be collected to examine the relationships between achievement goals, online community identification and collaborative reflection more deeply. Second, the state-of-the-art language understanding and feature extraction methods like RoBERTa, ALBERT, and XLNet can be considered in further research. In addition, implementing educational applications and evaluating the effects according to the findings of this study are promising, such as developing a co-reflection platform or a personalized feedback system (Xie, Chu, Hwang, & Wang, 2019), exploring the integrating of the co-reflection platform and teaching (Zou, Xie, Wang, & Kwan, 2020), and investigating the feedback of teachers and students (Hwang, Yang, & Wang, 2013).

## Acknowledgement

This work was supported by the Humanities and Social Sciences Planning Fund of the Ministry of Education, China (No. 18YJA880027).

## References

- Anderman, E. M. (2010). Reflections on Wittrock's generative model of learning: A Motivation perspective. *Educational Psychologist*, 45(1), 55-60.
- Bowskill, N. (2017). Sharedthinking: A Social identity approach to critical thinking. *Journal of Pedagogic Development*, 7(2), 37-46.
- Carvalho, A. M. (2009). Scoring functions for learning Bayesian networks. *INESC-ID Technical Reports*, 12. Retrieved from [http://www.lx.it.pt/~asmc/pub/talks/09-TA/ta\\_pres.pdf](http://www.lx.it.pt/~asmc/pub/talks/09-TA/ta_pres.pdf)
- Chang, Y., Hou, R. J., Wang, K., Cui, A. P., & Zhang, C. B. (2020). Effects of intrinsic and extrinsic motivation on social loafing in online travel communities. *Computers in Human Behavior*, 109, 106360. doi:10.1016/j.chb.2020.106360
- Chen, X., Xie, H., & Hwang, G. J. (2020). A Multi-perspective study on artificial intelligence in education: Grants, conferences, journals, software tools, institutions, and researchers. *Computers & Education: Artificial Intelligence*, 1, 100005. doi:10.1016/j.caeai.2020.100005
- Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 1, 100002. doi:10.1016/j.caeai.2020.100002
- Chiu, C. M., Huang, H. Y., Cheng, H. L., & Sun, P. C. (2015). Understanding online community citizenship behaviors through social support and social identity. *International Journal of Information Management*, 35(4), 504-519.
- Clarà, M., Kelly, N., Mauri, T., & Danaher, P. A. (2017). Can massive communities of teachers facilitate collaborative reflection? Fractal design as a possible answer. *Asia-Pacific Journal of Teacher Education*, 45(1), 86-98.
- Collin, S., & Karsenti, T. (2011). The Collective dimension of reflective practice: The How and why. *Reflective Practice*, 12(4), 569-581.
- Dalgarno, B., Reupert, A., & Bishop, A. (2015). Blogging while on professional placement: Explaining the diversity in student attitudes and engagement. *Technology, Pedagogy and Education*, 24(2), 189-209.
- David, A. P. (2014). Analysis of the separation of task-based and self-based achievement goals in a Philippine sample. *Psychological Studies*, 59(4), 365-373.
- Deci, E. L., & Ryan, R. M. (2012). Self-determination theory. In P. A. M. V. Lange, A. W. Kruglanski & E. T. Higgins (Eds.), *Handbook of Theories of Social Psychology* (Vol. 1, pp. 416-437). Thousand Oaks, CA: Sage.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. Retrieved from <https://arxiv.org/abs/1810.04805>
- Elliot, A. J., Murayama, K., & Pekrun, R. (2011). A 3×2 achievement goal model. *Journal of Educational Psychology*, 103(3), 632-648.
- Elliot, A. J., & Thrash, T. M. (2001). Achievement goals and the hierarchical model of achievement motivation. *Educational Psychology Review*, 13(2), 139-156.

- Elliott, E. S., & Dweck, C. S. (1988). Goals: An Approach to motivation and achievement. *Journal of Personality and Social Psychology*, 54(1), 5–12.
- Ergün, E., & Avcı, Ü. (2018). Knowledge sharing self-efficacy, motivation and sense of community as predictors of knowledge receiving and giving behaviors. *Educational Technology & Society*, 21(3), 60-73.
- González-Carvajal, S., & Garrido-Merchán, E. C. (2020). Comparing BERT against traditional machine learning text classification. Retrieved from <https://arxiv.org/abs/2005.13012>
- Granberg, C. (2010). Social software for reflective dialogue: Questions about reflection and dialogue in student teachers' blogs. *Technology, Pedagogy and Education*, 19(3), 345-360.
- Hagger, M. S., & Hamilton, K. (2018). Motivational predictors of students' participation in out-of-school learning activities and academic attainment in science: An Application of the trans-contextual model using Bayesian path analysis. *Learning and Individual Differences*, 67, 232-244. doi:10.1016/j.lindif.2018.09.002
- Heckerman, D. (1997). Bayesian networks for data mining. *Data Mining and Knowledge Discovery*, 1(1), 79-119.
- Heo, D., Anwar, S., & Menekse, M. (2018). The Relationship between engineering students' achievement goals, reflection behaviors, and learning outcomes. *International Journal of Engineering Education*, 34(5), 1634-1643.
- Hew, K. F., Hu, X., Qiao, C., & Tang, Y. (2020). What predicts student satisfaction with MOOCs: A Gradient boosting trees supervised machine learning and sentiment analysis approach. *Computers & Education*, 145, 103724. doi:10.1016/j.compedu.2019.103724
- Huang, C. Q., Han, Z. M., Li, M. X., Jong, M. S. Y., & Tsai, C. C. (2019). Investigating students' interaction patterns and dynamic learning sentiments in online discussions. *Computers & Education*, 140, 103589. doi:10.1016/j.compedu.2019.05.015
- Hulleman, C. S., Schrager, S. M., Bodmann, S. M., & Harackiewicz, J. M. (2010). A Meta-analytic review of achievement goal measures: Different labels for the same constructs or different constructs with similar labels? *Psychological Bulletin*, 136(3), 422-449.
- Hwang, G. J., Sung, H. Y., Chang, S. C., & Huang, X. C. (2020). A Fuzzy expert system-based adaptive learning approach to improving students' learning performances by considering affective and cognitive factors. *Computers & Education: Artificial Intelligence*, 1, 00003. doi:10.1016/j.caeai.2020.100003
- Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Computers & Education: Artificial Intelligence*, 1, 100001. doi:10.1016/j.caeai.2020.100001
- 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. doi:10.1016/j.compedu.2013.07.008
- Kalk, K., Luik, P., & Taimalu, M. (2019). The Characteristics of students, blog groups and blogging that predict reflection in blogs during teaching practice and induction year. *Teaching and Teacher Education*, 86, 102900. doi:10.1016/j.tate.2019.102900
- Krutka, D. G., Bergman, D. J., Flores, R., Mason, K., & Jack, A. R. (2014). Microblogging about teaching: Nurturing participatory cultures through collaborative online reflection with pre-service teachers. *Teaching and Teacher Education*, 40, 83-93. doi:10.1016/j.tate.2014.02.002
- Lau, S., Liem, A. D., & Nie, Y. (2008). Task-and self-related pathways to deep learning: The Mediating role of achievement goals, classroom attentiveness, and group participation. *British Journal of Educational Psychology*, 78(4), 639-662.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.
- Lei, C., & Chan, C. K. (2018). Developing metadiscourse through reflective assessment in knowledge building environments. *Computers & Education*, 126, 153-169. doi:10.1016/j.compedu.2018.07.006
- Lim, J. Y., & Lim, K. Y. (2020). Co-regulation in collaborative learning: Grounded in achievement goal theory. *International Journal of Educational Research*, 103, 101621. doi:10.1016/j.ijer.2020.101621
- Liu, Q., Zhang, S., Wang, Q., & Chen, W. (2017). Mining online discussion data for understanding teachers reflective thinking. *IEEE Transactions on Learning Technologies*, 11(2), 243-254.
- Mercier, E. M. (2017). The Influence of achievement goals on collaborative interactions and knowledge convergence. *Learning and Instruction*, 50, 31-43. doi:10.1016/j.learninstruc.2016.11.006
- Meyer, K. A., & Xu, Y. J. (2007). A Bayesian analysis of the institutional and individual factors influencing faculty technology use. *The Internet and Higher Education*, 10(3), 184-195.
- Novakovich, J. (2016). Fostering critical thinking and reflection through blog-mediated peer feedback. *Journal of Computer Assisted Learning*, 32(1), 16-30.

- Ozkan, Y. (2019). Reflectivity of pre-service language teachers echoed through blogs. *Kasetsart Journal of Social Sciences*, 40(1), 155-163.
- Payne, S. C., Youngcourt, S. S., & Beaubien, J. M. (2007). A Meta-analytic examination of the goal orientation nomological net. *Journal of Applied Psychology*, 92(1), 128-150.
- Qu, H., & Lee, H. (2011). Travelers' social identification and membership behaviors in online travel community. *Tourism Management*, 32(6), 1262-1270.
- Tajfel, H. (1978). The achievement of group identification. In H. Tajfel (Ed.), *Differentiation between social groups: Studies in the psychology of intergroup relations* (pp. 61-76). London, UK: Academic Press.
- Thijs, J., & Fleischmann, F. (2015). Student–teacher relationships and achievement goal orientations: Examining student perceptions in an ethnically diverse sample. *Learning and Individual Differences*, 42, 53-63. doi:10.1016/j.lindif.2015.08.014
- Tikhomirova, T. S., & Kochetkov, N. V. (2018). Relationship between learning motivation and reflection in undergraduate students. *Psychological Science and Education*, 23(6), 97-106.
- Ullmann, T. D. (2019). Automated analysis of reflection in writing: Validating machine learning approaches. *International Journal of Artificial Intelligence in Education*, 29(2), 217-257.
- Urdan, T., & Kaplan, A. (2020). The Origins, evolution, and future directions of achievement goal theory. *Contemporary Educational Psychology*, 61, 101862. doi:10.1016/j.cedpsych.2020.101862
- Van den Broeck, A., Ferris, D. L., Chang, C. H., & Rosen, C. C. (2016). A Review of self-determination theory's basic psychological needs at work. *Journal of Management*, 42(5), 1195-1229.
- Van den Kieboom, L. A. (2013). Examining the mathematical knowledge for teaching involved in pre-service teachers' reflections. *Teaching and Teacher Education*, 35, 146-156. doi:10.1016/j.tate.2013.06.009
- Vygotsky, L. S. (1978). *Mind in society: The Development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, 104, 106189. doi:10.1016/j.chb.2019.106189
- Wang, X., Mei, X., Huang, Q., Han, Z., & Huang, C. (2021). Fine-grained learning performance prediction via adaptive sparse self-attention networks. *Information Sciences*, 545, 223-240. doi:10.1016/j.ins.2020.08.017
- Won, S., Wolters, C. A., & Mueller, S. A. (2018). Sense of belonging and self-regulated learning: Testing achievement goals as mediators. *The Journal of Experimental Education*, 86(3), 402-418.
- Xie, H., Chu, H. C., Hwang, G. J., & Wang, C. C. (2019). Trends and development in technology-enhanced adaptive/personalized learning: A Systematic review of journal publications from 2007 to 2017. *Computers & Education*, 140, 103599. doi:10.1016/j.compedu.2019.103599
- Yilmaz, R. (2016). Knowledge sharing behaviors in e-learning community: Exploring the role of academic self-efficacy and sense of community. *Computers in Human Behavior*, 63, 373-382. doi:10.1016/j.chb.2016.05.055
- Yilmaz, F. G. K., & Keser, H. (2016). The Impact of reflective thinking activities in e-learning: A Critical review of the empirical research. *Computers & Education*, 95, 163-173. doi:10.1016/j.compedu.2016.01.006
- Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3), 55-75.
- Yukawa, J. (2006). Co-reflection in online learning: Collaborative critical thinking as narrative. *International Journal of Computer-Supported Collaborative Learning*, 1(2), 203-228.
- Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A Review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation*, 31(7), 1235-1270.
- Zhou, Y., Huang, C., Hu, Q., Zhu, J., & Tang, Y. (2018). Personalized learning full-path recommendation model based on LSTM neural networks. *Information Sciences*, 444, 135-152. doi:10.1016/j.ins.2018.02.053
- Zou, D., Xie, H., Wang, F. L., & Kwan, R. (2020). Flipped learning with Wikipedia in higher education. *Studies in Higher Education*, 45(5), 1026-1045.
- Zumbrunn, S., McKim, C., Buhs, E., & Hawley, L. R. (2014). Support, belonging, motivation, and engagement in the college classroom: A mixed method study. *Instructional Science*, 42(5), 661-684.

# STEM-based Artificial Intelligence Learning in General Education for Non-Engineering Undergraduate Students

Chun-Hung Lin<sup>1</sup>, Chih-Chang Yu<sup>2</sup>, Po-Kang Shih<sup>3</sup> and Leon Yufeng Wu<sup>4\*</sup>

<sup>1</sup>Center for Teacher Education, Chung Yuan Christian University, Taiwan // <sup>2</sup>Department of Information and Computer Engineering, Chung Yuan Christian University, Taiwan // <sup>3</sup>Center for General Education, Chung Yuan Christian University, Taiwan // <sup>4</sup>Graduate School of Education, Chung Yuan Christian University, Taiwan // chlin@cycu.edu.tw // ccyu@cycu.edu.tw // sportgun@cycu.edu.tw // leonwu@cycu.edu.tw

\*Corresponding author

**ABSTRACT:** This article describes STEM education with artificial intelligence (AI) learning, particularly for non-engineering undergraduate students. In the designed three-week learning activities, students were encouraged to put their ideas about AI into practice through two hands-on activities, utilizing a provided deep learning-based web service. This study designed pre-test and post-test surveys to investigate the performance of students in different aspects of AI. With 328 students involved in these learning activities, we discovered from the surveys that the proposed learning method can effectively improve AI literacy among non-engineering students. This study also found that students' AI literacy correlated significantly with their awareness of AI ethical issues and that the STEM-based AI curriculum increased the awareness of AI ethical issues among low-AI-literate learners. This article discusses the association between learning activities and different aspects of AI learning. The proposed method can be used by teachers who want to introduce AI knowledge into general education courses.

**Keywords:** Artificial intelligence, STEM education, General education, Non-engineering students, Artificial intelligence literacy

## 1. Introduction

Artificial intelligence (AI) has become part of the educational curriculum and educational services in modern society (Goel, 2017). For non-engineering students, it is important to learn the basic concepts of AI and to develop their understanding of AI and its application directions so that they can picture a future AI-enriched world. The learning process of AI education requires students to combine knowledge from different fields. Hence, incorporating AI learning into STEM education is worthwhile at this moment because STEM education focuses on interdisciplinary learning experiences. Previous studies have identified the trend of incorporating STEM integration into education to foster future citizenship in science (Li et al., 2020). Although a previous study indicated that pre-college math and science test scores and levels of confidence in other related quantitative skills (i.e., ACT math and science test scores and placement test scores) may be used to distinguish non-engineering students from others (Veenstra et al., 2008), educators have increasingly stressed an STEM-integration orientated learning infrastructure for non-engineering students (Nathan et al., 2013) because the interdisciplinary learning experiences reflecting science, technology, engineering, and mathematics are connecting the authentic world more than ever (Katehi et al., 2009). However, numerous studies have depicted negative pre-mindsets among non-engineering students toward learning in a pro-STEM environment (Hu et al., 2020). Non-engineering students often reveal that they are disconnected from the real world in a conventional learning environment. Owing to the nature of non-engineering students' training processes, the value and trends of STEM-based interdisciplinary learning are not always understood among non-engineering students (Lau et al., 2016; Lo et al., 2017).

For most non-engineering students, STEM-related courses are not their primary interests or requirements. Thus, in their learning paths in higher education, science, technology, engineering, and math are not the focus in their learning portfolios. In most cases, it is easier to reach these non-engineering students through STEM-related introductory courses in the general education curriculum. To gain a deeper understanding of what factors affect non-engineering students, this study aimed to understand how different students' backgrounds and characteristics affect their understanding of AI and awareness of AI ethical issues in the course.

### 1.1. Scientific introductory courses in general education

The core value of the university was holistic education, which included general education and intellectual education. However, with modernization, the goal of university education was repositioned to cultivate

professional and technical skills. General education can make up for the shortcomings of current education and improve students' creativity, comprehensive ability, judgment, critical ability, and cognitive skills, so that students can have cross-discipline cooperation ability and develop more mature personalities (Pan & Pan, 2005). General education allows students to realize the world from multiple perspectives and find the meaning and purpose of life through a culture-type course.

In addition to the humanities, science is an important part of general education (Kirk-Kuwaye & Sano-Franchini, 2015). Scientific introductory courses emphasize the "spirit of science" and "scientific literacy" for non-engineering students, as courses that focus solely on "knowledge" would be boring and would therefore reduce students' learning motivation (Pintrich, 1990). Discussing philosophy is sometimes too abstract and unreal; it is crucial in science learning that the design of instructional materials is relevant to authentic daily life (Abd-El-Khalick et al., 1998). Therefore, linking life experiences is essential when appropriating learning materials to achieve better scientific literacy (DeBoer, 2000). Allowing students to learn and understand the science applied in life can help them think about the meaning of science and stimulate their interest in learning (Glynn et al., 2005). For example, using augmented reality (AR) and virtual reality (VR) technologies to teach an astronomy course, students were motivated and encouraged with the assistance of technologies (Liou et al., 2017).

### **1.2. Instructional strategies for STEM education**

Previous studies have confirmed that STEM education has a significant impact on student learning outcomes, especially among Asian students (Wahono et al., 2020; Dong et al., 2020). STEM education can be implemented in various ways. Experiencing a successful STEM education depends not only on the teachers' beliefs, knowledge, and understandings but also on the adequate instruction of STEM concepts (McMullin & Reeve, 2014; Dong et al., 2020). Many past studies have attempted to incorporate problem-based learning strategies into STEM education and have found that they have a positive impact on students' learning outcomes (Sayary, Forawi, & Mansour, 2015; Wang, 2020). With a proper design, STEM education can be extended to ubiquitous learning (Wu et al., 2013). Following the above suggestions, this study adopts a problem-based learning strategy to design the learning activities. The challenge here is to apply STEM enactment to students who are unfamiliar with AI technology. Section 3 describes the details.

Hands-on scientific courses in general education can improve students' motivation and self-confidence in learning science and technology (Krupczak et al., 2005). STEM courses can be part of the university's general education courses that are designed to let students understand not only humanities, writing and literature, and history but also the sciences (including mathematics and technology). For college and university students not majoring in science or engineering, STEM courses can help them look for ways to solve problems and strengthen their educational experiences for future job opportunities (Enderson & Ritz, 2016).

### **1.3. Important scientific issues: AI education and literacy**

The content of teaching and the way of learning about scientific issues need to keep pace with the times (Huang, 2005). For example, technology education has become a basic learning content, and the rise of AI in recent years has been one of the most important scientific and technological issues (Cantú-Ortiz et al., 2020). AI had been developed rapidly and was widely used in different fields, such as manufacturing, economy, communications, transportation, medical care, and education (Pan, 2018).

Thinking about how to teach AI has become important because people's demand for AI applications has increased; however, it is not easy to design a proper AI course which matches students' expertise in the educational field. Allowing the application of AI technology to integrate closely with educational theory can help students obtain a more basic and comprehensive understanding of this topic (Chen et al., 2020). In addition to teaching knowledge, adding practical content to AI courses can increase students' learning motivation (Kostaris et al., 2017). Lin and colleagues (2021) discovered that intrinsic motivation has a significant influence on career motivation. Therefore, educators should foster students' intrinsic motivation and design appropriate instructional strategies so that students wish to strengthen their career motivation by pursuing AI-related knowledge.

In future education, students will learn not only knowledge but also literacy, which is a combination of knowledge, attitude, and skills. For example, scientific literacy is manifested in people's lifestyles and is the internalization of scientific knowledge and understanding of life (Maienschein, 1998). AI education can improve students' AI literacy. Moreover, AI education does not specifically refer to improving students' technical knowledge of, for example, programming, but rather their understanding of AI concepts and applications. The

application of AI is quite extensive, and improving students' AI literacy helps to strengthen their ability to cooperate and communicate with others so that students can recognize and solve problems (Konishi, 2016; Long & Magerko, 2020).

#### **1.4. Ethics as a Social Scientific Issue (SSI) element for an AI course**

As an important technology widely utilized in daily life, AI has greatly impacted people's lives in many ways. Hwang et al. (2020) documented the connotation of AI education from several angles, such as the development of learning models, implementation frameworks, and learning systems. As such, researchers ought to revisit existing educational theories, learning strategies, and methods to reflect this emerging knowledge in education.

In addition to the increasing stress of learning about basic knowledge of AI, ethical issues regarding the practice of AI technology are equally stressed in current education connotations. Due to various prejudices and algorithms that lack humanity, the abuse of AI violates human rights and inequalities. This violation is an ethical issue that will generally attract people's attention. Therefore, it is necessary to emphasize human-centered AI, enhance learners' awareness of ethical issues through education, and implement the moral teaching of AI for the practitioners (Goldsmith & Burton, 2017; Yang et al., 2021). The course design should not only help students understand the knowledge of AI but also emphasize the impact of AI technology on morality. In recent years, increasing research in the AI field has raised ethical issues. From 2016 to 2018, discussions of interchange, fairness, responsibility, and sustainability increased in AI academic papers (Jobin et al., 2019; Hagendorff, 2020). Research on AI literacy should discuss these issues.

#### **1.5. Purpose of the current study**

Since AI is an important scientific issue in this era, it has been regarded as a priority in higher education. For engineering students, AI is a kind of technology. However, for non-engineering students, AI is more likely to be a tool. Hence, the designed learning unit was placed in a general education course with participants who were all non-engineering students. This study investigates the AI literacy of non-engineering university students and identifies the differences in students' AI literacy before and after receiving related courses; the findings will serve as a reference for future curriculum development and revision. The study also determines the impact of the STEM-based course on learners' awareness of AI issues among learners with different AI literacy levels.

This study attempts to answer the following two questions:

- (1) Does the STEM-based AI course have an impact on the understanding of AI and AI literacy among students from different majors?
- (2) Do different levels of AI literacy have an impact on students' awareness of AI ethical issues?

## **2. Methodology**

### **2.1. Participants**

This study involved 328 non-engineering freshmen from various majors at a university in Taiwan. There were 40–65 students per course and 13 classes. In terms of gender distribution, 108 students were male (32.9%), and 220 students were female (67.1%). Of the students, 79 were from the Department of Accounting (24.1%), 71 were from the Department of Business Management (21.6%), 65 were from the Department of Information Management (19.8%), 41 were from the Department of Landscape Architecture (12.5%), 23 were from the Department of Applied Linguistics and Language Studies (7%), 22 were from the Department of Finance (6.7%), 16 were from the Department of International Trade (4.9%), 4 were from the Department of Teaching Chinese as a Second Language (1.6%), 2 were from the Department of Special Education (0.6%), 2 were from the Undergraduate Program in Social Design (0.6%), 2 were from Department of Financial And Economic Law (0.6%), and one was from the Department of Commercial Design (0.3%). The study was approved by the campus ethics committee, and all participants agreed to participate in the experiments.

## 2.2. Procedure

This study designed a three-week AI course as part of a regular 18-week general education course, Introduction to Science and Technology, at a university in northern Taiwan. The designed course consisted of lectures for the first week and hands-on exercises for the following two weeks (see Table 1). All participants had a two-hour activity each week.

To evaluate the potential contribution of the proposed AI literacy cultivation, a pre-test was administered before the course to survey learners' AI literacy, AI understanding, and awareness of AI ethics. After the pre-test survey, a lecture was given to establish a baseline of the learners' knowledge in week one. We then provided a series of instructions for hands-on activities in week two. The third week involved a small summative exercise requiring students to utilize the knowledge they had learned in week two to train an AI model that could recognize the "moving directions" and apply the model to a "motor-controlled car kit built on Raspberry Pi" (see Figures 1 and 2). After the activities were completed in week three, a post-test was applied to evaluate whether there was a learning effect on students' AI literacy, understanding, and awareness of AI ethics.

*Table 1. The STEM-based AI course unit design*

Weeks	Activities
Week 1	Pre-test (10 mins) Lecture (110 mins)
Week 2	Train an AI model (60 mins) Create an object recognition application (60 mins)
Week 3	Train an AI model that can recognize road signs (30 mins) Apply the model to the car kit (90 mins) Post-test (10 mins)

*Table 2. Connections between AI learning activities and knowledge points*

Item	Corresponding Learning Activities	Knowledge Point (AI Understanding Question Items)
1	Lecture in week 1	I think AI can generate new knowledge.
2	Activity in week 2	I think we must collect enough data to create a good AI model.
3	Activity in weeks 2 & 3	Programming language is required for designing AI applications.
4	Lecture in week 1	I think AI improves its accuracy by reducing certain errors.
5	Lecture in week 1	Deep learning is an AI technique.
6	Activity in weeks 2 & 3	I think the abilities of current AI models are limited.
7	Activity in weeks 2 & 3	I think the algorithm of designing an AI model is important.
8	Lecture in week 1 Activity in weeks 2 & 3	I think most existing AI models are task-specific.

Several training activities were conducted in each of the three weeks. Each proposed activity correlated highly with knowledge points in the survey items (See Table 2). In the first week, students were taught about several important topics in the AI field, including the history of AI and what the scientists are trying to achieve, the definition of supervised learning and unsupervised learning, applications in AI, and ethical issues within the development of AI. The three professional lecturers were from the Departments of Information and Computer Engineering, Information Management, and Electrical Engineering. The purpose of the first week of lectures was to provide students with a basic understanding of AI technology, including its purpose, achievable goals, and current bottlenecks.

Because engineering students lack knowledge of the programming language required to design an AI model, this study utilized a web service called Custom Vision, provided by Microsoft Azure. Custom Vision utilizes a deep learning technique called convolutional neural network (CNN) and provides a web-based interface for users to train their models by adopting transfer learning (Zhang et al., 2018). The provided interface hides the implementation details of creating an AI model, but allows flexibility in designing the problem. Users need only to upload an image dataset to complete an object recognition task. This tool is highly suitable for this study, which aims to teach students how to solve problems using AI techniques. In the training activity in week two, we used a public dataset containing 25,000 images of two types of objects. Each student was given a training dataset and a test dataset. The training dataset contained two classes (e.g., images of cats and dogs), and each class contained 500 images. The test dataset had 20 images. We asked the students to perform the following three experiments:



- (1) Use five images from each class to train the model and then test the accuracy using all test datasets.
- (2) Use 20 images from each class to train the model and then test the classification accuracy.
- (3) Increase the number of images to train the model until all images in the test dataset are correctly classified.

In this training activity, the students discovered that the trained model was highly inaccurate when using only 10 training images. However, the accuracy improved to almost 100% after using more than 100 images for training. This exercise gave them the knowledge that sufficient training data are needed to create a good AI model. In addition, students were asked to choose a picture containing neither of the two classes, feed this picture to the network, and observe the recognition result. The purpose of this exercise was to let students know that the AI model can output only what it knows. For the model to be able to distinguish between “true objects” and “none of the above,” it is necessary to provide additional images that do not contain any desired objects (called negative samples), set them into a category, and allow the model to learn them.

In the training activity in week three, students were separated into groups, and each group was provided with a motor-controlled car kit, as shown in Figure 1. The car kit was built using a Raspberry Pi with Raspbian OS so that it could execute programs. The car kits used in this study were equipped with USB cameras to capture images. We also provided two types of road signs (i.e., moving directions for the cars: a left-turn sign and a right-turn sign). In this training activity, the students were asked to design a model that could drive the wheels under different circumstances. More specifically, when the car kit “saw” a right-turn sign, it should turn right. Similarly, it should turn left when the car kit “saw” the left-turn sign. To accomplish this task, students had to first collect several road sign images, upload the images to the Custom Vision website, and then train a model. In addition, students had to consider how to react when the car did not “see” any road signs. For example, if the car stopped at a crossroad, it had to keep waiting until it saw a road sign. At that moment, it would try to detect whether there was a road sign in front of it or not. That is, they had to collect images that represented negative samples, and this practice was related to the activity they did in week two. After finishing the design, the teaching assistants in the class helped the students deploy the model on the car kit. This step was slightly technical, so we intentionally avoided having students do it on their own. Students could determine if there was a problem with the model they had designed by how it behaved on the car kit. If the model did not perform well enough—for example, the car was unable to recognize the road sign correctly—they were encouraged to collect the data again and train a better model.

In this training activity, students realized that they had to design a proper algorithm so that the car kit could respond correctly. We asked the students to engage in this exercise in groups. Because the algorithm they designed may have contained flaws, the teachers needed to guide them in revising their algorithm through discussions. Even if the algorithm were designed well, the car kit might sometimes not have reacted correctly due to the wrong recognition results on road signs. In addition, they could not directly take the model they trained in week two and tackle the problems they encountered in week three. While these activities contain knowledge related to understanding items 6–8 in Table 2, it is worth pointing out that the current AI model was a purpose-specific model, not a generic one. Figure 3 summarizes the relationships between the knowledge points and training activities.



*Figure 1. Car kit “motor-controlled car kit built on Raspberry Pi” used in this study*



*Figure 2.* Actual teaching scenario in this study. A student was holding one road sign of moving direction and training the motor-controlled car kit built on Raspberry Pi

In summary, the three-week course began by introducing the basic concepts of AI, presenting several applications that would allow students to think about the development of AI, and introducing ways to solve problems with current AI models. In the first week of interaction, the students had some questions about current AI technologies. To validate their questions, the students had to train their AI models to solve certain problems over the next two weeks. From these experiments, they understood several basic concepts of AI. First, the training of AI models is a data-learning mechanism, which means that we must provide enough data to make the model accurate. Second, the current AI model is task-specific; in other words, it does not know how to solve the given problem. Therefore, it is necessary for humans to design an appropriate algorithm to solve problems with the help of the AI model. Finally, even if properly designed, AI models still have their limitations. It is still possible for an AI model to give wrong decisions. How to deal with these anomalies is an important task for humans.

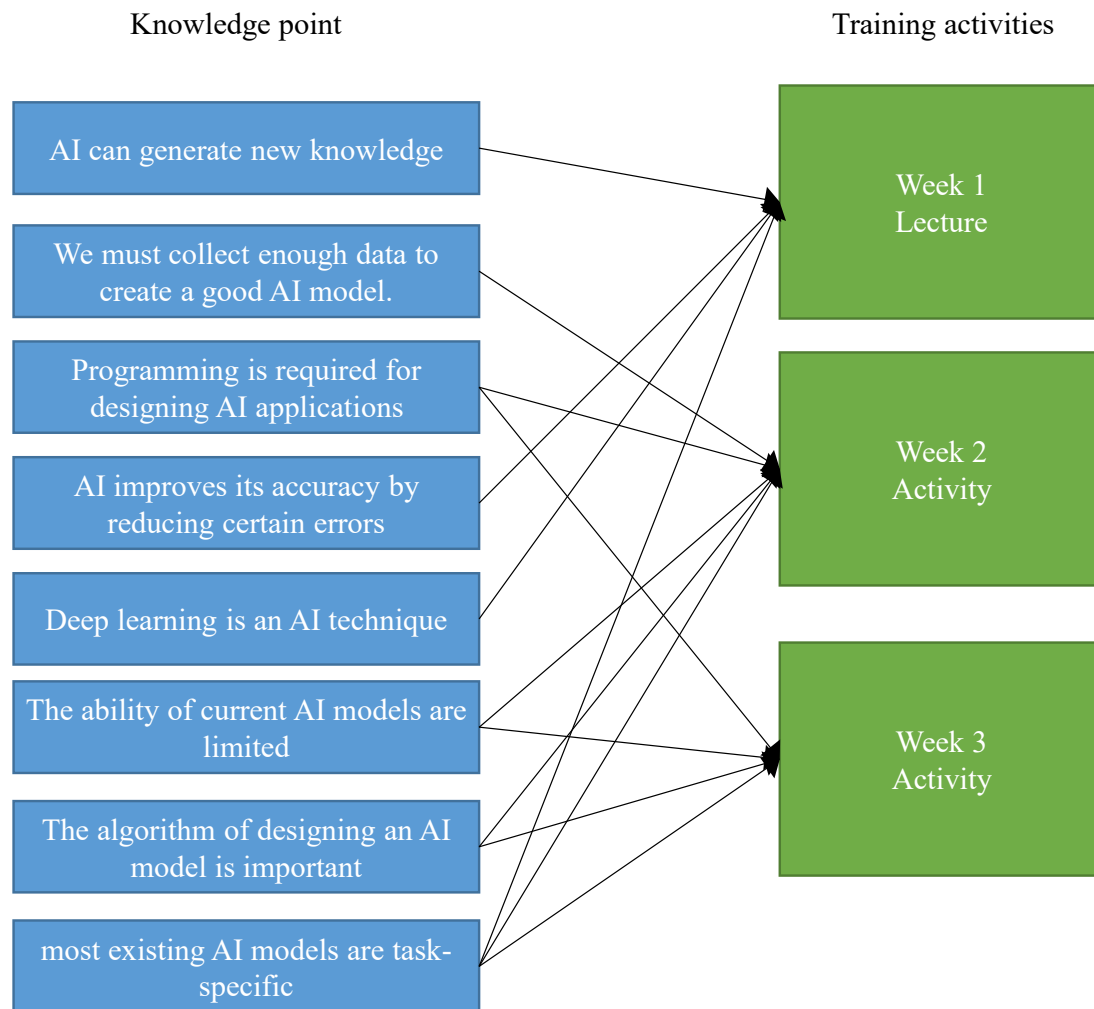


Figure 3. Connections between all knowledge points and the designed three-week lecture and activities

## 2.3. Instrument

### 2.3.1. AI literacy scale (AI literacy)

This study adapted an AI literacy scale developed by Lin et al. (2020) to evaluate learners' AI literacy. This scale is designed in the form of a Likert-style five-point scale with 1 corresponding to "strongly disagree" and 5 corresponding to "strongly agree." To understand the important factors of AI, this study applied factor analysis to construct validity. The result of the KMO value was .945, and the significant value of Bartlett's spherical test was .000, suggesting that the dataset was suitable for factor analysis and could explain up to 65.29% of variance. Finally, two important aspects were extracted: (1) teamwork (four items) and (2) attitude toward AI (eight items). The overall internal consistency reliability (Cronbach's alpha) was 0.943, suggesting that the scale maintained good reliability. As illustrated in Section 3.2, the pre-test survey was administered before the first learning activity, and the post-test was administered at the end of the third week.

### 2.3.2. AI understanding scale (AI understanding)

Eight question items were designed with the revision comments of three experts in relevant fields to estimate the learners' levels of AI understanding after the course. During the course, these question items also served as knowledge points to better align with the design of the lessons and the learning activities. The AI understanding survey was also designed in the form of a Likert-style five-point scale with 1 corresponding to "strongly disagree" and 5 corresponding to "strongly agree." To justify, AI is an ongoing area of science, just like other areas of science still in the process of continuing development. Often, some scientific statements merely describe the current state of development and might not always be true in the future. Therefore, we placed this set of

questions to estimate students' levels of AI understanding in our experimental design. Further, the instructors of these courses also employed these questions as discussion topics during the courses.

### 2.3.3. AI ethics awareness scale (AI ethics)

To understand learners' awareness of the ethical issues of AI, this study developed an AI awareness scale with references to the findings of Jobin et al. (2019) and Hagendorff (2020). The scale was developed using a five-point Likert-type scale, with 1 corresponding to "strongly disagree" and 5 corresponding to "strongly agree." The scale contained 15 questions on four dimensions: Transparency (1,2,3,4), Responsibility (12,13,14), Justice (7,8,9,10), and Benefit (16,17,18,19). The reliability of the overall scale was higher than 0.7, indicating that the scale had good reliability.

## 3. Results and discussion

In this section, several analyses were conducted to respond to the research questions raised in this study.

### 3.1. Research question 1: Does the STEM-based AI course have an impact on the understanding of AI among students from different majors?

To estimate learners' levels of AI understanding after the STEM-based AI courses, a repeated *t*-test analysis was applied in this study. The comparisons between the pre- and post-tests (see Table 3) showed that the score of students' AI understanding (mean value) increased from 4.02 to 4.13, and the standard deviation was .60 and .62 respectively. The *t*-value was 2.99 ( $p = .003 < .01$ ), indicating a significant difference between the pre-test and the post-test scores. The results showed that non-engineering students' levels of AI understanding improved significantly after the course. Hence, we can infer that the present AI course can help enhance students' understanding of AI. Furthermore, it was pointed out that hands-on activities in STEM courses are an important element that can effectively enhance students' active learning and increase their learning effectiveness (Yannier et al., 2020; Mater et al., 2020). The experimental results in this study also matched this viewpoint, showing that students' understanding of AI improved through hands-on activities.

Table 3. Results of the repeated *t*-test analysis on students' understanding of AI

	<i>N</i>	Pre-test		Post-test		<i>t</i>
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Understand of AI	328	4.02	.60	4.13	.62	2.99**

Note. \*\* $p < .01$ .

To determine whether there were differences in students' understanding of AI among students from different majors, we used students' pre-test scores of AI understanding as a covariate and students' post-test scores of AI understanding as the dependent variable, and we applied ANCOVA. No significant difference was found in students' understanding of AI from different majors (see Table 4). After examining the performance of students' pre-test and post-test scores, we observed that students' post-test scores were higher than their pre-test scores but not at a significant level (see Table 5).

Table 4. Results of the analysis of covariance on students' understanding of AI across different majors

Majors	<i>N</i>	Mean	<i>SD</i>	Adjusted mean	<i>F</i>
Accounting	79	4.06	.58	4.10	.67
Business Management	71	4.16	.62	4.15	
Information Management	65	4.27	.60	4.22	
Landscape Architecture	41	4.07	.73	4.07	
Applied Linguistics and Language Studies	23	4.13	.43	4.10	
Finance	22	4.15	.61	4.16	
International Trade	16	4.03	.74	4.06	
Teaching Chinese as a Second Language	4	3.81	.75	3.93	
Special Education	2	4.38	.18	3.93	
Social Design	2	3.75	1.06	3.92	
Financial and Economic Law	2	4.00	.00	3.85	
Commercial Design	1	3.00	-	3.34	

Table 5. results of the repeated t-test analysis on students' understanding of AI across different majors

Majors	N	Pre-test		Post-test		t
		M	SD	M	SD	
Accounting	79	3.94	.59	4.06	.58	1.67
Business Management	71	4.05	.51	4.16	.62	1.47
Information Management	65	4.14	.55	4.27	.60	1.74
Landscape Architecture	41	4.02	.81	4.07	.73	.39
Applied Linguistics and Language Studies	23	4.09	.54	4.13	.43	.50
Finance	22	3.99	.61	4.15	.61	1.20
International Trade	16	3.64	.68	4.03	.74	.40
Teaching Chinese as a Second Language	4	3.75	.65	3.81	.75	.24
Special Education	2	3.62	.53	4.38	.18	1.50
Social Design	2	3.65	.88	3.75	1.06	1.00
Financial And Economic Law	2	4.38	.53	4.00	.00	1.00

### 3.2. Research question 2: Does the STEM-based AI course have an impact on the AI literacy of students from different majors?

Table 6 depicts the effect of STEM-based AI courses on overall students' AI literacy. The analysis results show that after the STEM-based AI course, students' performance in the two dimensions of AI literacy—attitude toward AI and teamwork—improved significantly, indicating that non-engineering students' AI literacy can be positively enhanced through the proposed STEM-based AI course.

Table 6. Results of the repeated t-tests on students' AI literacy

AI literacy	N	Pre-test		Post-test		t
		M	SD	M	SD	
Attitude toward AI	328	4.07	.65	4.14	.69	2.02*
Teamwork	328	3.42	.71	3.83	.73	10.10***

Note. \* $p < .05$ ; \*\*\* $p < .001$ .

As mentioned above, this study found that the hands-on activities in the present AI courses improved students' understanding of AI. Furthermore, this study found that combining hands-on activities with group work helped enhance non-engineering students' perceptions of AI issues and strengthen their awareness of interdisciplinary teamwork. In the process of completing tasks related to AI through teamwork, learners can have the opportunity to realize that cooperation is an important channel for completing tasks related to AI, and this awareness is an important part of AI literacy. This finding echoes those of a previous study (Hurson et al., 2011).

To determine whether there was a difference in students' learning performance in AI literacy among students from different majors, we used students' pre-test scores of AI literacy as a covariate and students' post-test scores of AI literacy as the dependent variable, and we applied ANCOVA. The results showed that there was no significant difference in students' AI literacy from different majors (Table 7). After comparing the performance of students' pre-test and post-test scores, we observed that students' post-test scores were higher than their pre-test scores. Moreover, the results of the repeated t-test analysis showed that students from the Department of Accounting, Business Management, Information Management, Landscape Architecture, Applied Linguistics and Language Studies, and Finance showed significant changes in AI literacy. By contrast, students from the Department of International Trade (IT), Teaching Chinese as a Second Language (TCSL), Special Education (SE), Social Design (SD), and Financial and Economic Law (FEL) did not reach significant differences in AI literacy (Table 8). It is thought that a desired outcome was not observed among some majors due to unbalanced participant data, as the results revealed that those majors with relatively more participants benefited from the course. Therefore, we conjecture that, overall, the STEM-based course may have an impact on participants from different majors.

AI ethics are important for everyone due to the maturity of AI technology. Kocanjer and Kadoić (2016) recommended a method to raise students' ethical awareness by organizing workshops or debates on topics of ethics. Takahara and Kajiwara (2013) adopted debates in engineering ethics classes to improve students' communication skills. From their research, we can conclude that debate is a good activity for raising students' ethical awareness. This study also designed some question items to estimate the ethical awareness of students. Table 9 tabulates the correlation between students' AI literacy and their awareness of AI ethical issues. From an analysis of the results, we can see that the scores for awareness of AI ethical issues positively correlated with AI

literacy, which shows that a correlation exists between different AI literacies and students' perceptions of AI ethics.

*Table 7. Results of the analysis of covariance on students' AI literacy across different majors*

Majors	<i>N</i>	Mean	<i>SD</i>	Adjusted mean	<i>F</i>
Accounting	79	3.90	.63	3.93	.55
Business Management	71	3.97	.61	3.96	
Information Management	65	4.21	.61	4.10	
Landscape Architecture	41	4.02	.72	4.01	
Applied Linguistics and Language Studies	23	3.83	.67	3.90	
Finance	22	4.03	.70	4.04	
International Trade	16	3.95	.68	4.01	
Teaching Chinese as a Second Language	4	3.84	1.12	4.16	
Special Education	2	3.69	.09	3.93	
Social Design	2	3.81	1.15	4.02	
Financial and Economic Law	2	2.94	.09	3.54	
Commercial Design	1	3.00	-	3.85	

*Table 8. Results of the repeated *t*-test analysis on students' AI literacy across different majors*

Majors	<i>N</i>	Pre-test		Post-test		<i>t</i>
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Accounting	79	3.70	.57	3.90	.63	2.70**
Business Management	71	3.76	.56	3.97	.61	2.57*
Information Management	65	3.93	.56	4.21	.61	3.39**
Landscape Architecture	41	3.76	.56	4.02	.72	2.97**
Applied Linguistics and Language Studies	23	3.63	.56	3.83	.67	2.39*
Finance	22	3.73	.50	4.03	.70	3.04**
International Trade	16	3.63	.63	3.95	.68	1.40
Teaching Chinese as a Second Language	4	3.19	.65	3.84	1.12	1.50
Special Education	2	3.31	.27	3.69	.09	1.66
Social Design	2	3.38	.71	3.81	1.15	1.37
Financial and Economic Law	2	2.68	.27	2.94	.09	2.00

Note. \* $p < .05$ ; \*\* $p < .001$ .

*Table 9. Correlation between students' AI literacy and awareness of AI ethical issues*

	Transparency	Benefit	Justice	Responsibility	Cognition	Awareness	Teamwork
Transparency	-						
Benefit	.80***	-					
Justice	.85***	.86***	-				
Responsibility	.45***	.88***	.86***	-			
Cognition	.38**	.41***	.37***	.37***	-		
Awareness	.46***	.44***	.43***	.45***	.60***	-	
Teamwork	.30***	.31***	.31***	.25***	.47***	.49***	-

Note. \*\* $p < .01$ ; \*\*\* $p < .001$ .

To further confirm the relationship between AI literacy and perceptions of AI ethical issues, this study analyzed students' perceptions of AI ethical issues. By utilizing ANOVA, the data were categorized into different levels of AI literacy. The overall result between the pre-test and post-test was not significant. As mentioned above, fostering ethical awareness in students requires in-depth interactions and discussions over a long period. Because the proposed course design comprised only six hours over three weeks, there was not enough time to organize enough interactions that might inspire students' ethical awareness via the learning activities. However, this study adopted the teacher-directed strategy to bring some ethical cases into the discussions in-depth. For example, teachers asked students about the ethical issues of autonomous vehicles: "How will the machine react if it has to make choices, like the classical trolley problem? Or say, how should it react?" Students elaborated on or debated their thoughts in class with their peers. To verify the performance of this design, we divided the responses from the students on AI literacy into two levels: high and low. Then, we examined each dimension by their levels of AI literacy. Tables 10 and 11 present the results of the analysis.

Table 10. Effects of AI literacy on students' awareness of AI ethical issues

Dimensions	<i>N</i>	<i>df</i>	<i>MS</i>	<i>F</i>	Group	<i>M</i> ( <i>SD</i> )
Transparency	328	1	15.39	41.76***	High	4.53(.54)
					Low	4.10(.68)
Benefit	328	1	15.24	39.05***	High	4.53(.57)
					Low	4.09(.68)
Justice	328	1	15.41	39.69***	High	4.53(.57)
					Low	4.09(.68)
Responsibility	328	1	14.16	35.90***	High	4.56(.58)
					Low	4.14(.68)

Note. \*\*\* $p < .001$ .

Table 11. Changes in Awareness of AI Ethical Issues among Students with Different AI Literacies

Dimensions	Group	<i>N</i>	Pre-test		Post-test		<i>t</i>
			<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Transparency	High	179	4.60	.47	4.53	.54	1.49
	Low	149	3.99	.62	4.10	.68	2.10*
Benefit	High	179	4.57	.47	4.53	.57	1.16
	Low	149	3.96	.66	4.09	.68	2.70**
Justice	High	179	4.57	.48	4.53	.58	0.56
	Low	149	3.98	.65	4.09	.76	2.43*
Responsibility	High	179	4.65	.47	4.56	.58	2.03*
	Low	149	4.04	.69	4.14	.68	1.93

Note. \* $p < .05$ ; \*\* $p < .01$ .

The results suggest that the higher the level of AI literacy students possess, the higher the level of their awareness of AI ethics. This characteristic is reflected in the four dimensions of ethics: transparency, benefit, justice, and responsibility. A further comparison of the averages of AI ethics shows that students with high AI literacy have significantly higher levels of transparency, benefit, justice, and responsibility than students with low AI literacy (see Table 10). In addition, students with lower levels of AI literacy benefited more significantly from the course on all four dimensions than those with higher levels of AI literacy (see Table 11).

The ethical issues of AI have received increasing attention in recent years, and this study found that the performance of students' AI literacy in the context of STEM-based courses was significantly and positively correlated with students' awareness of AI ethical issues. The results in Table 11 revealed that bringing case discussions of ethical issues into STEM-based curricula helped to increase low AI-literate learners' awareness of AI ethical issues. However, the proposed type of discussion was not effective in increasing the awareness of AI ethical issues among high AI literate students. This finding is likely because teacher-directed case-based instruction is effective in increasing students' basic concepts of ethical issues but not in enhancing learners' higher levels of ethical issues (Takahara & Kajiwar, 2013). Learners with higher ethical literacy require more sophisticated teaching methods and activities, such as group debates and case studies.

#### 4. Conclusions and future work

This study proposes a set of STEM-based course modules in the form of lectures, case discussions, and hands-on activities for students with non-engineering backgrounds. These course lessons were developed with a supporting previous framework of AI literacy. Several findings during the analysis showed that the course effectively improved students' AI literacy (i.e., perceptions toward teamwork in an AI-enriched environment and AI adoption) among non-engineering students. The students' AI literacy was correlated with their awareness of AI ethics, and increments occurred in the levels of awareness of AI ethics among learners with low AI literacy. On the contrary, we found that students with high literacy could experience less or limited awareness of ethical issues. For the high-literate students, what might have occurred during the course at various points was not discovered in the current study. A possible future direction could be to customize some more challenging hands-on activities for higher AI-literate learners to see if their levels of awareness of ethical issues could be developed during their teamwork. Designing instructions for different levels of certain perceptions toward core course objectives is reasonable. Furthermore, the experience or ability to define and discuss the problems encountered with their AI car kit was an important learning objective in relation to STEM learning.



In this study, we discussed the effects of a STEM-based AI course on students' understanding of AI and examined its effects on students' awareness of ethical issues in AI. A positive correlation was found between students' AI literacy and their awareness of AI ethical issues. We found that learners with high AI literacy showed a higher awareness of AI ethical issues. In AI education, instructors usually place great emphasis on students' engagement in AI tasks, motivation to learn AI-related content, and learning performance in AI-related topics. Whether AI literacy has the same impact on these dimensions is a valuable direction for future research. Through this kind of study, we can deepen our understanding of the relationship between AI literacy and students' AI learning.

General education is an appropriate way to cultivate students' literacy. For non-engineering students, general education is a medium through which to expose them to important scientific issues, such as AI. We designed a three-week AI course, merged AI literacy into the course, and obtained positive results. In other words, the designed lessons expand the scope and purpose of scientific introductory courses in general education by including the field of AI. This study provides suggestions based on empirical evidence for future STEM-based AI instructional designs.

## Acknowledgement

This study was partially funded by the Ministry of Science and Technology and the Ministry of Education, Taiwan (R.O.C.). under grant no. MOST 109-2221-E-033-033-, MOST 108-2511-H-033-003-MY2 and PGE1090426.

## References

- Abd-El-Khalick, F., Bell R. L., & Lederman N. G. (1998). The Nature of science and instructional practice: Making the unnatural natural. *Science Education*, 82, 417-436.
- Burton, E., Goldsmith, J., Koenig, S., Kuipers, B., Mattei, N., & Walsh, T. (2017). Ethical considerations in Artificial Intelligence courses, *AI Magazine*, 38(2), 22-34.
- Cantú-Ortiz, F. J., Sánchez, N. G., Garrido, L., Terashima-Marin, H., & Brena, R. F. (2020). An Artificial intelligence educational strategy for the digital transformation. *International Journal on Interactive Design and Manufacturing*, 14, 1195-1209.
- Chen, X. L., Xie, H. R., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 1, 100002. doi: 10.1016/j.caeai.2020.100002
- DeBoer, G. E. (2000). Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. *Journal of Research in Science Teaching*, 37(6), 582-601.
- Dong, Y., Wang, J., & Yang, Y. (2020). Understanding intrinsic challenges to STEM instructional practices for Chinese teachers based on their beliefs and knowledge base. *IJ STEM Ed*, 7, 47. doi:10.1186/s40594-020-00245-0
- Enderson, M. C., & Ritz J. (2016). STEM in general education: Does mathematics competence influence course selection. *The Journal of Technology Studies*, 42(1), 30-40.
- Glynn, S. M., Aultman L. P., & Owens A. M. (2005). Motivation to learn in general education programs. *The Journal of General Education*, 54(2), 150-170.
- Goel, A. K. (2017). AI education for the world. *AI Magazine*, 38(2), 3-4.
- Goldsmith, J., & Burton, E. (2017). Why teaching ethics to AI practitioners is important. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), 4836-4840.
- Hagendorff, T. (2020). The Ethics of AI ethics: An Evaluation of guidelines. *Minds and Machines*, 30, 99-120.
- Hu, C. C., Yeh, H. C., & Chen, N. S. (2020). Enhancing STEM competence by making electronic musical pencil for non-engineering students. *Computers & Education*, 150, 103840. doi:10.1016/j.compedu.2020.103840
- Huang, C. J. (2005). A Study of using science news as general education program teaching materials. *Nanhua General Education Research*, 2(2), 59-83.
- Hwang, G. J., Xie, H. R., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 1, 100001. doi:10.1016/j.caeai.2020.100001



- Hurson, A. R., Sedigh, S., Miller, L., & Shirazi, B. (2011). Enriching STEM education through personalization and teaching collaboration. In *Proceedings of the 7th IEEE International Conference on Pervasive Computing and Communications Workshops*. doi:10.1109/PERCOMW.2011.5766949
- Jobin A., Ienca, M., & Vayena, E. (2019). The Global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1, 389–399.
- Katehi, L., Pearson, G., & Feder, M. (2009). *Engineering in K-12 education: Understanding the status and improving the prospects*. Washington, WA: National Research Council.
- Kirk-Kuwaye, M., & Sano-Franchini D. (2015). Why do I have to take this course? How academic advisers can help students find personal meaning and purpose in general education. *The Journal of General Education*, 64(2), 99-105.
- Kocanjer, D., & Kadoić, N. (2016). Raising students' awareness about ethical behavior. In *Proceedings of the 4th Global Virtual Conference* (pp. 88-93). doi:10.18638/gv.2016.4.1.764
- Konishi, Y. (2016). What is needed for AI literacy? *RIETI Special Series: Priorities for the Japanese Economy in 2016*. Retrieved from [https://www.rieti.go.jp/en/columns/s16\\_0014.html](https://www.rieti.go.jp/en/columns/s16_0014.html)
- Kostaris, C., Sergis, S., Sampson, D. G., Giannakos, M. N., & Pelliccione, L. (2017). Investigating the potential of the flipped classroom model in K-12 ICT teaching and learning: An Action research study. *Educational Technology & Society*, 20(1), 261-273.
- Krupczak, J. J., Vanderstoep, S., Wessman, L., Makowski, N., Otto, C. A. & Dyk, K. (2005). Work in progress - Case study of a technological literacy and non-majors engineering course. In *Proceedings Frontiers in Education 35th Annual Conference* (SIJ-36). doi:10.1109/FIE.2005.1612208
- Lau, C., Lo, K., Chan, S., & Ngai, G. (2016). From zero to one: Integrating engineering and non-engineering students in a service-learning engineering project. In *Proceedings of the Second International Conference on Service-Learning (ICSL 2016)* (pp. 186-191).
- Li, Y., Wang, K., Xiao, Y., & Froyd, J. E. (2020). Research and trends in STEM education: A Systematic review of journal publications. *International Journal of STEM Education*, 7(1), 11. doi:10.1186/s40594-020-00207-6
- Liou, H. H., Yang, S. J., Chen, S. Y., & Tarn, W. (2017). The Influences of the 2D image-based augmented reality and virtual reality on student learning. *Educational Technology & Society*, 20(3), 110-121.
- Lin, C. H., Wu, L. Y., Wang, W. C., Wu, P. L., & Cheng, S. Y. (2020, February). *Development and validation of an instrument for AI-Literacy*. Paper presented at the 3rd Eurasian Conference on Educational Innovation (ECEI 2020), Hanoi, Vietnam.
- Lin, P. Y., Chai, C. S., Jong, M. S. Y., Dai, Y., Guo, Y. M. & Qin, J. J. (2021). Modeling the structural relationship among primary students' motivation to learn artificial intelligence. *Computers and Education: Artificial Intelligence*, 2, 100006. doi:10.1016/j.caeai.2020.100006
- Lo, K. W. K., Lau, C. K., Chan, S. C. F., & Ngai, G. (2017). When non-engineering students work on an international service-learning engineering project - A Case study. In *Proceedings of the 2017 IEEE Global Humanitarian Technology Conference (GHTC)*. doi:10.1109/GHTC.2017.8239292
- Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). doi:10.1145/3313831.3376727
- Maienschein, J. (1998). Scientific Literacy. *Science*, 281(5379), 917-917.
- Mater, N. R., Hussein, M. J. H., Salha, S. H., Draidi, F. R., Shaqour, A. Z., Qatanani, N., & Affouneh, S. (2020). The Effect of the integration of STEM on critical thinking and technology acceptance model. *Educational Studies*. doi:10.1080/03055698.2020.1793736
- McMullin, K., & Reeve, E. (2014). Identifying perceptions that contribute to the development of successful project lead the way pre-engineering programs in Utah. *Journal of Technology Education*, 26(1), 22–46.
- Nathan, M. J., Srisurichan, R., Walkington, C., Wolfgram, M., Williams, C., & Alibali, M. W. (2013). Building cohesion across representations: A Mechanism for STEM integration. *Journal of Engineering Education*, 102(1), 77-116.
- Ng, O. L., Shi, L., & Ting, F. (2020). Exploring differences in primary students' geometry learning outcomes in two technology-enhanced environments: Dynamic geometry and 3D printing. *International Journal of STEM Education*, 7, 50. doi:10.1186/s40594-020-00244-1
- Pan, J. D., & Pan, H. M. (2005). Accomplishments of the holistic education idea: An Example from Chung Yuan Christian University. *Chung Yuan Journal*, 33(2), 237-251.
- Pan, Y. H. (2018). 2018 special issue on artificial intelligence 2.0: theories and applications. *Frontiers of Information Technology & Electronic*, 19(1), 1-2.

- Pintrich, P. R., & de Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, 82(1), 33–40.
- Sayary, A. M. A., Forawi, S. A., & Mansour, N. (2015). STEM education and problem-based learning. In *The Routledge International Handbook of Research on Teaching Thinking* (pp. 357-368). New York, NY: Routledge.
- Takahara, K., & Kajiwara, T. (2013). Engineering ethics education on the basis of continuous education to improve communication ability. *Electrical Engineering in Japan*, 183(3), 1–8.
- Veenstra, C. P., Dey, E. L., & Herrin, G. D. (2008). Is Modeling of freshman engineering success different from modeling of non-engineering success? *Journal of Engineering Education*, 97(4), 467–479.
- Wahono, B., Lin, P. L., & Chang, C. Y. (2020). Evidence of STEM enactment effectiveness in Asian student learning outcomes. *International Journal of STEM Education*, 7, 36. doi:10.1186/s40594-020-00236-1
- Wang, Y. (2020). Integrating games, e-books and AR techniques to support project-based science learning. *Educational Technology & Society*, 23(3), 53–67.
- Wu, P. H., Hwang, G. J., & Tsai, W. H. (2013). An Expert system-based context-aware ubiquitous learning approach for conducting science learning activities. *Educational Technology and Society*, 16, 217–230.
- Yang, S. J. H., Ogata, H., Matsui, T. & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 2, 100008. doi: 10.1016/j.caeai.2021.100008
- Yannier, N., Hudson, S.E., & Koedinger, K.R. (2020). Active learning is about more than hands-on: A Mixed-reality AI system to support STEM education. *International Journal of Artificial Intelligence in Education*, 30, 74–96.
- Zhang K., Cheng H. D., & Zhang B. (2018). Unified approach to pavement crack and sealed crack detection using preclassification based on transfer learning. *Journal of Computing in Civil Engineering*, 32(2), 04018001. doi:10.1061/(ASCE)CP.1943-5487.0000736

# Progress, Challenges and Countermeasures of Adaptive Learning: A Systematic Review

Fengying Li<sup>1</sup>, Yifeng He<sup>2</sup> and Qingshui Xue<sup>3\*</sup>

<sup>1</sup>School of Continuing Education, Shanghai Jiaotong University, Shanghai, China // <sup>2</sup>Division for Development of Liberal Arts, Shanghai Jiaotong University, Shanghai, China // <sup>3</sup>School of Computer Science and Information Engineering, Shanghai Institute of Technology, Shanghai, China // [fyli@sjtu.edu.cn](mailto:fyli@sjtu.edu.cn) // [heyifeng0712@sjtu.edu.cn](mailto:heyifeng0712@sjtu.edu.cn) // [xue-qsh@sit.edu.cn](mailto:xue-qsh@sit.edu.cn)

\*Corresponding author

**ABSTRACT:** With the deep application of artificial intelligence and big data in education, adaptive learning has become a new research hotspot in online education. Based on the systematic review of the connotation and research progress of adaptive learning, a new definition of adaptive learning is given. By literature analysis, this paper points out the challenges faced by adaptive learning research, such as the lack of cognition of brain and technology, the bottleneck of the model of emotion domain, the separation of education and technology, the security of data management and the risk of privacy leakage. These challenges can be summarized into two aspects: one is mechanical issues, the other is safety issues. Different from traditional research perspectives, the paper opens a new research window, and puts forward countermeasures from the perspectives of cognitive principles, zone of proximal development theory in technology, breakthrough in the emotional domain model, learning data management and privacy security. In view of the centralization of learning data management nowadays, the concept of code chain and the decentralized management mode based on code chain are proposed. Different from the traditional adaptive learning recommendation technology, a new adaptive learning pulling model is proposed.

**Keywords:** Adaptive learning, Learning recommendation, Learning pulling, Code chain technology, Data security

## 1. Introduction

In recent years, “learner-centered” - based individualized education and learning has become a new trend in the development of education in the world. Education departments of lots of countries have formulated action plans in response. Singapore has implemented the plan that each student has a learning terminal to support students’ personalized learning (Lan, 2015). The “Vision 2020” published by the UK government sets out the relevant issues that need to be addressed in personalized learning (Li, 2008). In South Korea, the Ministry of Education, Science and Technology issued the implementation plan of promoting intelligent education strategy, to comprehensively carry out intelligent education and implement personalized teaching and learning (Piao, 2012). The U.S. Department of Education released the report “Promoting Teaching and Learning through Educational Data Mining and Learning Analysis” to really realize personalized learning with the help of big data (Xu, Wang, Liu, & Zhang, 2013). Ten-year Development Plan for Educational Informatization in China (2011-2020) points out that “an information-based environment should be built to provide personalized learning services for each student” (The Ministry of Education of China, 2012). However, most current online education platforms only share learning resources, including teachers, courses and other hardwares and softwares, to learners, but fail to provide targeted learning support at a specific time. In addition, for a large number of learners, it is difficult for teachers to achieve effective interaction with learners. Therefore, learners’ individual needs cannot be really met.

However, the personalization of learning can be achieved using various methods that have been made available by the rapid development of Information Communication Technology (ICT) (Dawson, Heathcote, & Poole, 2010). Adaptive or personalized learning has become possible by implementing intelligent learning systems, integrating learners’ preferences, analyzing individual learning data, and so on (Xie, Chu, Hwang, & Wang, 2019). Adaptive learning can achieve the requirements of personalized learning. According to different learners’ learning styles, learning levels and cognitive abilities, it can provide targeted services, such as learning content and path recommendation and intelligent tutoring, and provide personalized learning support for learners.

Adaptive learning is an emerging development that has been mentioned in The NMC Horizon Report (Higher Education Edition) since 2004 by The New Media Consortium. It is recognized as a major advance in higher education for two consecutive years in 2015 and 2016. The World Economic Forum (2020) released a report entitled “The School of the Future: Defining a New Education Model for the Fourth Industrial Revolution,” which proposed a global framework of “Education 4.0,” namely eight key characteristics of learning content and

experience. The seventh and eighth key features especially emphasize the importance of personalized learning and autonomous learning. This fully indicates that adaptive learning has become an important proposition and a new teaching paradigm in the development of education in lots of countries (Xie, Chu, Hwang, & Wang, 2019), and the research on adaptive learning has become a major topic in the field of education science. Xie et al. (2019) points out that technology-enabled adaptive or personalized learning has been a popular and important research direction in the field of educational technology.

Theoretically, with the continuous emergence of new technologies such as human-computer interaction in online education, sentiment analysis, big data and intelligent robots, and the deep integration of artificial intelligence technology and education, adaptive learning enables students to break through the limitations of regions and time. According to their own needs, learners can independently control the learning progress, improve the learning efficiency, and can achieve open, sharing and ideal learning state. But in the reality of teaching and learning, it does not produce the expected effects. On the one hand, the adaptive learning system is still immature and faces many difficulties and challenges, such as the separation between education and technology, the transmission bottleneck in the emotional field, the security of learning data management and the disclosure of learner privacy. On the other hand, the homogenization of adaptive learning is serious (Chen, 2003). Different from the previous research perspective, first of all, we define the concept of adaptive learning from a new perspective. Then, through literature analysis, we systematically elaborate the research progress, and try to find out the specific problems in the study of adaptive learning. Next, through deep analysis, the main causes of these problems are found out. Finally, innovative theory and coping strategies are put forward. To sum up, the research questions are:

- What are the problems and challenges faced by adaptive learning research at present?
- What are the causes of the problems?
- What are our strategies in the face of challenges?

## 2. Literature review

The essence of adaptive learning is scaleable personalized learning, which is closely related to autonomous learning, personalized learning and individualized learning, and its concept is easily confused. At the same time, with the development of information technology, its research connotation has also changed a lot. According to the technical level, adaptive learning is divided into three kinds: artificial learning, computer programming and artificial intelligence. Modern adaptive learning, as a product of artificial intelligence and “Internet +” education era, has been regarded as a new educational technology innovation (iResearch, 2018). Therefore, it is necessary to reorganize and define again from four categories: concept, system, model and technology.

### 2.1. Concepts of adaptive learning

There is no unified view on the concept of adaptive learning in academic circles. It is generally accepted that adaptive learning refers to the learning mode that provides corresponding learning environment, examples or fields for learning, and through the discovery and summary of learners themselves in learning, finally forms theories and can solve problems independently. Peter Brusilovsky (1996) from the University of Pittsburgh proposed that adaptive learning is based on individual differences in learners’ knowledge background, learning attitude, learning style, learning ability and other aspects. Zhu (1997) put forward the “conditional construction-optimization theory” of adaptive learning, and systematically elaborated the information processing process in which people acquire knowledge and skills through example learning. From the perspective of “teaching,” Zhao, Xu, and Long (2015) believed that adaptive learning meant that teachers used adaptive learning systems as teaching aids to collect and analyze data, prepare lessons, understand the learning state, evaluate, and timely adjust the teaching content to meet the changing learning needs of students. Wang and Wang (2014) argued from the perspective of “learning” that adaptive learning was to obtain learning content, way and path suitable for oneself through adaptive learning system. From the perspective of “learning tool support,” Chandrasekaran et al. (1992) and Corbett and Anderson (1994) believed that adaptive learning was to model a knowledge system by combining the knowledge level of students with the intelligent tutoring systems based on knowledge and adaptive learning systems, and then recommend a knowledge construction route to them.

It can be seen that the early concept of adaptive learning is mostly from the perspective of traditional pedagogy, without highlighting the influence of intelligence and intelligent technology. With the continuous integration of intelligent technology, the research of “AI+ adaptive learning” has become a new proposition in international research, and adaptive learning has also been endowed with new meanings.

We believe that adaptive learning is an autonomous, intelligent, technology-driven and individualized learning approach guided by teaching and learning theories. Accordingly, adaptive learning system is an online learning environment or learning support/service system that integrates the concept of adaptive learning into it. The connotation of adaptive learning integrated with AI or intelligent technology is changing from “self-adaptation” to “intelligent adaptation,” with new attributes different from the traditional meaning: autonomy, intelligence, individualization and adaptability.

Adaptive learning has two core words with iconic characteristics: “self” and “adaptation.”

“Self” is first manifested as the learner’s self-consciousness and autonomy, which emphasizes the student-centered autonomous learning. This kind of learning, different from passive learning, rote learning or indoctrination learning, is close to the visceral “meaningful learning” advocated by Ausubel (1960). “Self” is also embodied in the aspect of intelligence, that is, according to learners’ self-characteristics, to automatically guide learners to deepen their cognition. It makes automatic recording of learning process, learning behaviors and learning results. According to the learning process data, it can automatically judge, automatically associate learning resource, automatically evaluate and automatically adjust learning strategies and learning behaviors. Hwang et al. (2020) pointed out AI-supported learning systems can simulate human intelligence to reason, judge or predict, not only to provide personalized guidance, support or feedback to students, but also to help teachers or decision-makers make decisions.

“Adaptation” is firstly manifested as individualization, that is, learners can independently choose their own learning methods and learning contents that meet their own development needs. In addition, learners can set their own learning schedule and learn in the most comfortable way, which fully demonstrates the “autonomous learning concept” advocated in the global framework of Education 4.0. “Adaptation” is also reflected in the dynamic mutual adaptation and constant adjustment between learners and learning environment, the difficulty of learning content, learning partners (including teachers) and learning technology, so as to find the balance point among various elements in adaptive learning. The more adaptable you are, the more comfortable the learning process is and the more efficient the learning process is, which is also different from the previous interpretation of “adaptive.”

## 2.2. Adaptive learning system

Adaptive learning system, as an important carrier to support adaptive learning, has a close relationship with information technology. It can be said that the development of adaptive learning system has gone through six stages, including program teaching machine, computer-aided teaching, intelligent teaching system, intelligent agent teaching system, intelligent hypermedia teaching system and adaptive intelligent learning system. See Table 1 for details.

*Table 1. The six stages of the development of adaptive learning system (Tang, 2020; Li, Dong, & Tang, 2020)*

Stages	Time	Whether or not smart	Man-machine interaction mode	Learning system expression mechanism	Learning path	Theoretical basis	Instructional design
Program Instruction (PI)	1920s - 1960s	NO	Linear input/output	Knowledge showing	Preinstall	Behaviourism	Teaching-centered
Computer-Aided Instruction (CAI)	1970s	NO	Linear input/output	Knowledge showing	Preinstall	Behaviourism	Teaching-centered
Intelligent Teaching System (ITS)	1980s	AI	Multidimensional representation computing	Knowledge+ Induction	Preinstall	Behaviourism	Teaching-centered
Intelligent Agent Teaching System (Agent)	1990-1996	AI+ mass data	Perception	Knowledge+Data+ Computation+ Deduction	Preinstall +Recommendation	Cognitivism	Change from teaching-centered to learning-centered
Intelligent Hypermedia Teaching System	1997-2011	AI+ Mass data/ Big data	Perception +Lower cognition	Knowledge+ Mass data/ Big data+ Computing + Deduction	Preinstall +Recommendation	Cognitivism	Learning-centered

(AEHS)	2011-2017	AI+ Big data	Perception +Lower cognition	Knowledge+ Big data +Cloud computing+ Deduction	Preinstall +Recommen dation		
Adaptive Intelligent Learning System	2017-	AI+ Big data	Perception +Advanced cognition	Knowledge+ Big data +Cloud computing+ Deduction (Decision)	Preinstall +Recommen dation	Cognitivism	Learning-centered

In the 1950s, the programmed teaching proposed by Skinner can be called the germination of adaptive learning system; Computer Aided Instruction (CAI), which appeared in the 1970s, can be regarded as the prototype of the adaptive learning system. Pask in the UK developed the adaptive teaching machine using Computer, which is regarded as the primitive ancestor of CAI. Intelligent Tutoring Systems (ITS) emerged in the 1980s, and is often referred to as the earlier adaptive learning system, whose basic framework was proposed by Hartley and Sleeman (1973). In the 1990s, virtual reality (VR) and Agent technology were applied to ITS. Intelligent Agent teaching system, also known as intelligent student self-study software system, appeared. At the end of the 20th century and the beginning of the 21st century, the combination of artificial intelligence and Hypermedia technology has produced a new learning System, namely Adaptive Hypermedia System (AHS). In 1996, AEHS (Adaptive Educational Hypermedia System), developed by Professor Brusilovsky from the University of Pittsburgh in the United States, was called the first real Adaptive learning System (Brusilovsky, 1996). In recent years, with the continuous emergence of new technologies such as human-computer interaction, sentiment analysis and big data processing in online education, and the deep integration of artificial intelligence with educational science and psychology, the research on adaptive learning is deepening. Various adaptive learning systems have emerged, such as Knewton in the US, Knowre in South Korea, Smart Sparrow in Australia, online teacher training platform Declara, Cogbooks in the UK, Ape Test Bank, Classba Education and Homework Help in China.

Through the above six stages, it is not difficult to find that the research and application of adaptive learning systems generally present the following evolutionary trajectories. They are from intelligent teaching system to adaptive learning system, from non-intelligence to intelligence, from perception to cognition, from low-level cognition to advanced cognition including preliminary consciousness, from behaviorism to cognitivism, from preset learning path to learning recommendation, and from “teaching” as the center to “learning” as the center. By analysis of learning data, adaptive learning system adjusts learning content, knowledge assessment methods and knowledge sequence in real time, so as to meet learners’ personalized needs.

### 2.3. Adaptive learning model

Over the years, the research and evolution of adaptive learning model generally presents a continuous deepening and expanding from system model to module component model.

**ITS model** is regarded as the predecessor of adaptive learning. Hartley and Sleeman (1973), a British scholar from the University of Leeds, proposed the basic framework of ITS. This framework includes three basic models: (1) Domain knowledge, namely Expert Model; (2) Learner knowledge, i.e., Student Model; (3) Teaching strategy knowledge, namely the Tutor Model. The framework theory of ITS constituted by these three models has become the classical theory guiding the design and development of ITS.

**AHS model** was proposed by Peter Brusilovsky (1996) from the University of Pittsburgh, USA, based on the framework model of ITS. It is the first general model of Adaptive Hypermedia system, also known as AEHS (Adaptive Educational Hypermedia Systems). The model is divided into four core components: domain model, pedagogy model, student model and interface module. The four components are connected through the adaptive engine, which makes personalized resource recommendation to students through the personalized mechanism.

Brusilovsky has done a series of fruitful work in the aspects of adaptive learning theory and technology, and is regarded as the pioneer in this field. In addition to the general model of AEHS, he also proposed the intelligent guidance system ITEM /IP (Brusilovsky, 1992), and the adaptive learning system such as InterBook (Brusilovsky, Eklund, & Schwarz, 1998), EIM-Art (Weber & Brusilovsky, 2001), Knowledge Sea (Brusilovsky & Rizzo, 2003) and Annotat Ed (Farzan & Brusilovsky, 2008).

At the same time, based on Brusilovsky’s general model of adaptive learning (AEHS), extensive and in-depth studies have been carried out all over the world. Wolf of RMIT University in Melbourne had designed and

developed an adaptive learning environment using Java programming language - iWeaver, which uses the Dunn Learning Style Model (Wolf, 2003). Papanikolaou et al. (2003) from the University of Athens designed and developed a personalized education hypermedia system INspire, which generates course content according to learners' cognitive level and learning style. Alrifai et al. (2012) from the University of Hannover in Germany studied the user and domain model of adaptive learning system. Eindhoven University of Technology in the Netherlands developed an open source adaptive hypermedia system Aha!, which modified the user model and added new functions (AHA!, 2020). Wang, Zhao and Wei (2019) designed Mindolm, an open learner model, in the form of mind mapping visualization.

With the deepening of related researches, the functions of adaptive learning system model and component model become more and more rich. The system model develops from linear guidance to nonlinear guidance and from one-way broadcast to two-way interaction. The knowledge content of domain knowledge model develops from coarse granularity to fine granularity, the learner model develops from previous knowledge state analysis of students to learner style and emotion analysis, etc.

In our opinion, although many different models have been proposed, most of the researches on adaptive learning models are still based on AEHS model, which has not broken away from the traditional research pattern and has not achieved breakthrough progress.

## **2.4. Key technologies and algorithms of adaptive learning**

Different adaptive learning systems may have big differences in their function realization and content display. Generally speaking, there are three main ways to realize adaptive learning: adaptive content selection, adaptive navigation support, and adaptive content presentation (Brusilovsky, 2012; Romeroc & Zafraa, 2009). Nowadays, the new generation of adaptive learning system breaks the limitation of the traditional intelligent learning system that all students have the same learning path, and can create a customized learning content and learning path according to the learner's own state, namely learning data. And optimized learning programs are recommended to learners and personalized learning guidance is provided to learners that is different from others. The realization of this function mainly comes from the key technology and algorithms adopted by the system.

### **2.4.1. Key technologies**

The key technologies to truly realize personalized learning demand and learning pushing function mainly include data mining, learning analysis, machine learning, knowledge mapping, cognitive expert consultant, learning recommendation, edge computing, virtual reality, etc. The most commonly used adaptive learning techniques include Web application mining and text mining, semantic web ontology technology, fuzzy logic, etc. Romeroc and Zafraa (2009) integrated a specific Web mining tool and recommendation engine into AHA! The system helps teachers carry out the whole Web mining process, in the AHA! The system provides the most appropriate link page for students. Vesin et al. (2012) developed Protus2.0, an intelligent teaching system for learning programming languages, based on semantic web ontology technology, to create learner ontology, domain knowledge ontology, learning task ontology and teaching strategy ontology, and designed adaptive rules for reasoning to achieve personalized teaching. Chang et al. (2009) proposed a classification mechanism based on learners' Learning styles, optimized the K-Nearest Neighbor (KNN) classification algorithm, and combined it with GA (Gene Algorithm) algorithm and applied it in the open learning Management System, which could accurately and efficiently determine learners' Learning styles. Chrys Af Iadi and Virvou (2012) uses Kirkpatrick model and hierarchical evaluation method to evaluate the knowledge level of students, and uses fuzzy logic technology to define and update the knowledge level of students for the evaluation of ITS C language programming.

### **2.4.2. Algorithms**

Xu Kun (2020) combined with several adaptive learning platforms such as Kenton, Assissment and VIPKid, summed up three basic algorithms of adaptive learning: (1) Bayesian knowledge tracing. When tracking learners' mastery of knowledge points, Bayesian inference algorithm is used. (2) Bayesian network. Its basic structure is directed acyclic graph by analyzing, mining and modeling various association of learning data, so as to infer the path of students' learning evolution. These two Bayesian algorithms are collectively called probabilistic graph modeling. (3) Some technologies and methods in the field of educational measurement, such as Item response theory and Learning space theory. This algorithm can accurately locate the current knowledge level and learning

state of students for learning diagnosis and recommendation. In practice, these three basic algorithms are often expanded or combined to meet the specific needs.

## **2.5. Other aspects of adaptive learning research**

In recent years, the development of artificial intelligence (AI) has affected all areas of human life. The educational application of artificial intelligence has been widely concerned. AIED (Artificial Intelligence in Education) has been identified as the main research focus in the field of computer and Education (Chen, Xie, Zou, & Hwang, 2020; Hwang, Xie, Wah, & Gašević, 2020). AI-supported learning systems can simulate human intelligence to reason, judge or predict, not only to provide personalized guidance, support or feedback to students, but also to help teachers or decision-makers make decisions. Hwang, Xie, Wah, and Gašević (2020) proposed a framework to show the considerations of implementing AIED in different learning and teaching settings. The structure can help guide researchers with both computers and education backgrounds in conducting AIED studies. Chen, Xie, Zou, and Hwang (2020) evaluated definitions of AIED from broad and narrow perspectives and clarified the relationship among AIED, Educational Data Mining, Computer-Based Education, and Learning Analytics. Chen, Xie, and Hwang (2020) presented multiple perspectives on the development of AIED, and provided an overview of AIED for its further development and implementation. Zou & Xie (2018) developed a system based on Nation and Webb's checklist for technique feature analysis. This system recommends personalized word learning tasks based on the technique feature analysis scores of different tasks and user models. Based on human-computer collaboration, Li et al. (2019) proposed the construction method of knowledge graph in adaptive learning system, and took "artificial intelligence" discipline as an example to preliminarily verify the construction method.

In addition, from Interbook (Brusilovsky, 1998) in 1996, which focused on the personalized learning behavior of learners, to Cogbooks in 2015, which emphasized diversified analysis based on the personalized needs of learners, it can be seen that the adaptive learning system aims to continuously improve the learning process of students, provide interactive instructions in an automatic way, and provide learning support for learners anytime and anywhere (Walkington, 2013). Qiu, Zhao and Liu (2008) established the ontology of user model with text editor in 2008, and formed the database of user model. Jiang, Zhao and Wang (2011) adopted ontology technology to design the reference specification for establishing user model and knowledge model. Liu (2011) proposed the method of constructing the domain model and the corresponding design strategy of the process based on semantic network. In the study of mathematics, Ven et al. (2017) designed a tablet computer game, which can effectively help students improve their arithmetic ability of addition and subtraction. Stein (2019) emphasized that the biggest obstacle to personalized learning at present was the development of pedagogical theories to guide adaptive learning systems (Cui & Xu, 2019).

In the face of the current intelligent era, "digital generation" learners are increasingly pursuing diversified, personalized and comfortable learning needs, learning styles and learning scenarios, etc., and their requirements for learning analysis, learning evaluation and learning recommendation based on adaptive learning are also getting higher and higher. Therefore, we need to continue to explore new adaptive learning systems and applications.

## **3. Method**

The purpose of this study is to analyze and summarize the research trend and existing problems of adaptive learning in the world in recent years, and find out the path and method to solve the problems.

### **3.1. Data source**

The data of statistical analysis are mainly from the library of Shanghai Jiao Tong University, and the data collection is comprehensive. In the library of Shanghai Jiao Tong University, the collections are rich, and the quantity and quality of its electronic resources are among the best in China. Through the Shanghai Jiao Tong University Library - the entrance to the databases, CNKI and Scopus are selected, including full text of journals, full text of important newspapers, full text of important conference papers, full text of doctoral dissertations, full text of master's dissertations and other sub-databases.



### 3.2. Research method

The research includes academic literature retrieval, selection of retrieval results and analysis of sample data. Based on the literature database mentioned above, using “adaptive learning” as the key word, the retrieval scope was from 1971 to 2020, and the retrieval time was October 30, 2020. A total of 8,688 related literatures were retrieved, including 5,607 foreign literatures and 3,081 Chinese literatures. For those with repeated contents, the paper with the most complete data was selected after being judged to be the same study. A total of 7,880 papers in Chinese and English were detected after excluding unrelated papers, removing the duplicates and removing the literatures with unclear information sources and incomplete data. Because the library classifies “adaptive education,” “intelligent education,” “personalized learning” and other related concepts into “adaptive” learning category automatically, in order to comprehensively and accurately retrieve the required documents and avoid the exclusion of some relevant documents by the “precise” retrieval mode, this paper does not adopt the advanced precise retrieval mode with more retrieval conditions. Since the purpose of the analysis is to deeply understand the overall research trend of adaptive learning and the current problems encountered in the research of adaptive learning, the research mainly focuses on and analyzes the literature state and research problems in the recent 6 years from 2015 to 2020. Therefore, along with decades of research results, combined with the author’s research, reflection on the development and trend of adaptive learning has both support and guarantee.

### 3.3. Coding

The qualitative data coding method is used to process the data. Firstly, open coding is carried out for the researched problems, and all the problems related to adaptive learning are extracted. Then, spindle coding is made based on open coding. Finally, according to the researched questions, “core categories” are found out to complete the selection coding.

## 4. Results

### 4.1. Research trends in adaptive learning

By the literature quantity analysis, the research trend of “Adaptive Learning,” namely the trend chart of academic attention, is obtained, as shown in Figure 1. It can be found that the research trend of adaptive learning is on the rise on the whole. Early international attention on adaptive learning began before 1971, and the number of literatures increased year by year. From 2008 to 2019, the number of literatures showed an obvious upward trend of fluctuation. Especially from 2015 (454 articles) to 2019 (836 articles), the attention of the past five years has risen sharply, and the peak is reached in 2019. In 2020 (736), there is a slight decrease compared with 2019, and the difference is negligible. It has two reasons. One is the cause of the epidemic, and the other is incomplete data. This shows that adaptive learning has become a research hotspot in the academic field in recent years.

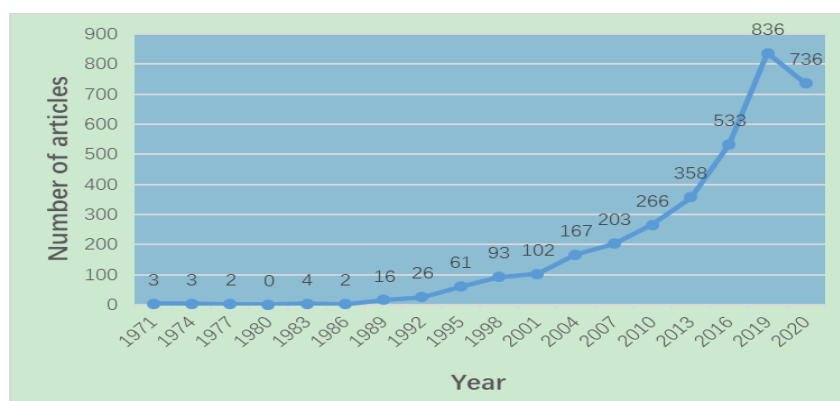


Figure 1. Academic attention on adaptive learning

Based on CNKI platform, with the theme of “adaptive learning research,” we used the measurement visualization analysis software inside CNKI platform to carry out the keyword co-occurrence network analysis and obtained the keyword co-occurrence network graph of self-adaptive learning research, as shown in Figure 2. It is found that the frequency of keywords such as neural network, adaptive learning system, personalized learning, adaptive control, machine learning and some algorithms is high, which indicates that the relationship between modern adaptive learning and artificial intelligence is very close.

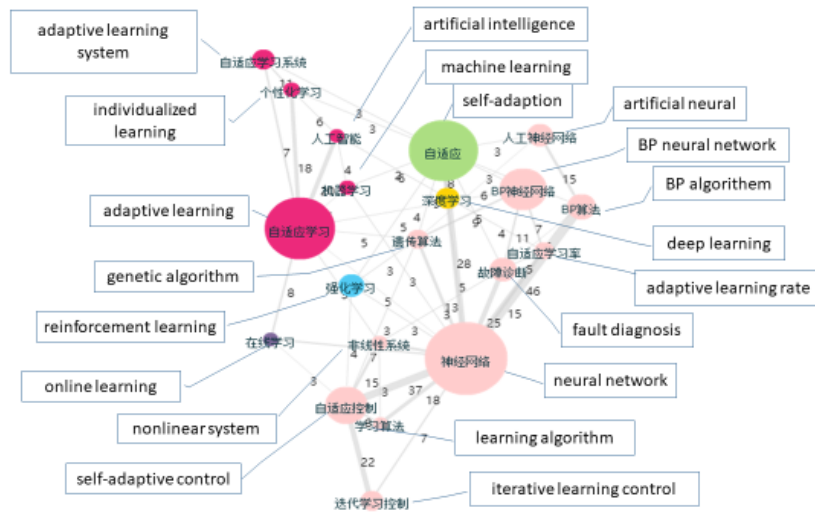


Figure 2. Adaptive learning studies co-occurrence networks of keywords

#### 4.2. Challenges in adaptive learning research

Through literature analysis, at present, the research on adaptive learning at home and abroad is undergoing a process of development and evolution from coarse to fine, from whole to module, and from theory to demonstration (Wu & Chen, 2018). From the microscopic point of view, from 2015 to 2020, there are 209 “problem research” literatures. It shows that there are still many difficulties and challenges in the study of adaptive learning. For example, there are more theoretical studies, less empirical studies, and less mature system platforms (Chen, Xie, Zou, & Hwang, 2020). There is a gap between theoretical research and practical application, and the practicability of research results is poor. (Xu & Wang, 2011; Chen, Xie, Zou, & Hwang, 2020). The data analysis results are as follows (Xie, Chu, Hwang, & Wang, 2019).

Table 2. The topic distribution of adaptive learning literature research “problem research”

Themes	Frequency	Proportion ( % )
Concept, understanding, policy	35	7
Technical problems	111	21
Disciplinary fragmentation	62	12
Emotional modeling	45	8
Data management	58	11
Privacy disclosure	42	8
Classification of knowledge points	37	9
More theoretical studies, less empirical studies	120	22
Others	22	4

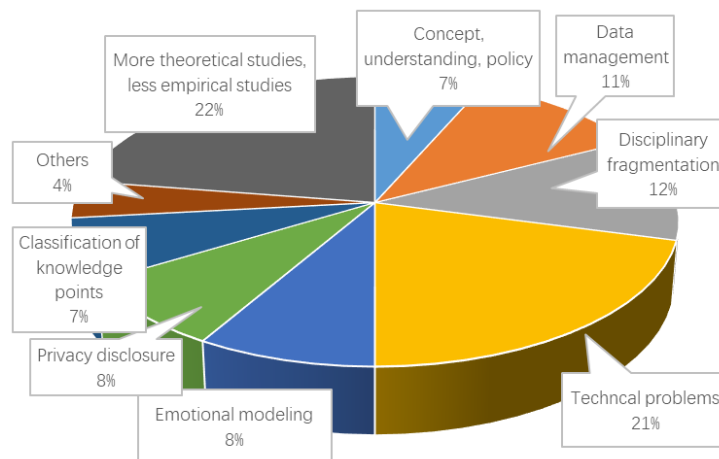


Figure 3. The topic distribution of adaptive learning research literature “problem research”

Outstanding performance in these problems: vague concept and unclear understanding (35), poor flexibility or technical issues (111), disciplinary fragmentation (62), modeling bottlenecks of emotion domain (45), learning data management (58), privacy security (42), broad classification of knowledge points (37), much theory but little practice (120) and others (22). The distribution is shown in Table 2 and Figure 3.

#### **4.2.1. Cognitive bottleneck**

Current research on adaptive learning lacks systematic and holistic cognition of learning, which is embodied in the following aspects: adaptive learning system is “clumsy,” not so “intelligent,” and cannot “follow one’s inclinations” to realize our learning ideas.

Due to the limitation of people’s cognition of the brain (Wu, 2020), the current artificial intelligence technology cannot meet the requirements of adaptive learning system. Although people have a clear understanding of the information transmission and processing principle of some neural circuits in the brain, as well as the mechanism of primary sensory function, the cognition of the global information processing, coding and learning principle of the brain is still very limited. The mathematical principles and computational models of information processing in the brain are still unclear (Wu & Pan, 2020). Therefore, the adaptive learning technology based on “artificial intelligence” which simulates brain function cannot achieve full intelligence, individuation and adaptability just like interpersonal communication.

Another misconception is the “technological omnipotence theory,” which holds that artificial intelligence is powerful enough to meet the requirements of adaptive learning systems (Nichols, 2020). Artificial intelligence has far more storage and computing power than the brain. For example, AlphaGo had beaten the human champion. However, training people is not the same as training machines. At the same time, in the face of huge courses and different knowledge systems, adaptive systems and platforms are difficult to cover all aspects. The narrow scope of knowledge contained by adaptive learning system will make adaptive learning lose part of its advantages in the development of future education world (Nichols, 2020). In addition, in the traditional adaptive learning mode, the logic of jump between different learning contents is linear and single. Even if students have mastered a certain content, they still need to spend time to learn it. What’s more, students can’t get immediate feedback or help when they have problems. Learning is a complex and implicit process, and simple computer programming is difficult to achieve good results.

#### **4.2.2. Disciplinary fragmentation**

A relatively perfect adaptive learning system often needs the close combination of theoretical guidance and technical implementation, and requires the cooperation of experts in many fields to complete. The realization of educational theory and method needs technology, which needs the guidance of educational thought. The reality is that there is a split between disciplines, particularly between educational science and computer science. The technical realization experts of adaptive learning system are mostly experts in the field of computer. Due to the lack of educational theory and personalized learning theory, it is difficult to design a learning system that conforms to the teaching law and learning law and is suitable for the personalized development. However, due to the lack of professional computer technology, scholars in the field of education are also unable to convert the concept of adaptive learning design into products (Xu & Wang, 2011; Wu & Chen, 2018). Hinton believed the key to overcoming the limitations of artificial intelligence was to build a bridge between computer science and biology (Somers, 2017).

#### **4.2.3. Lack of learner emotion modeling**

Learning is a complex and hidden process, people do not have the storage and computing power of a computer, but have seven emotions and six sensory pleasures, with complex physical and psychological performance, and these performance will have a complex impact on the learning experience. However, the traditional human-computer interaction is mechanical and difficult to meet the emotional needs of learners. At present, deep learning has been used to study emotion classification, but its achievements in natural language processing are not obvious, especially in the field of adaptive learning, there is no breakthrough. Adaptive learning should consider learners’ starting ability, learning style and emotional state, etc. However, the current system cannot understand learners’ emotions and cannot truly realize adaptive learning (Hu & Chen, 2018). Therefore, it is necessary to establish an adaptive learning algorithm and model that can fully understand the brain thinking and psychological emotions to perform human tasks. However, this emotion modeling process based on artificial intelligence is quite difficult and lacking (Cui & Xu, 2019).

#### **4.2.4. Flaws of centralization management of learning data**

The current adaptive learning systems, such as Knewton and Knowre, mostly adopt the centralized server management mode (Huang, Liu & Xue, 2020), which has three obvious deficiencies. First, this centralized service model is easy to be manipulated by others, which easily leads to the disclosure and attack of privacy and important data in the process of data analysis. Second, a large amount of learning data is unique to some institutions, which is easy to form Matthew Effect, resulting in difficulties in data collection and sharing, while adaptive learning analysis requires massive data and data sharing among different institutions and platforms. Third, centralized data interaction and management affect learning efficiency.

#### **4.2.5. Data security and privacy breaches**

Personalized learning is an important feature of adaptive learning. In order to achieve personalized and adaptive needs, the server of the adaptive learning system should collect and analyze the personal information of learners, such as learning interest, starting point level, learning style and emotional state. Adaptive learning evaluation also needs to collect learners' learning performance, learning process and the types of resources used by learners (Hu & Chen, 2018). After data collection, these learning data need to be analyzed before individualized recommendation of learning content and learning partners can be made. Otherwise, the learning process cannot be automatically customized, nor can personalized services be provided. In the process of learning data analysis, the privacy of learners, the security of learning data and the right to use will be involved (Cui & Xu, 2019).

In addition, with the development of artificial intelligence, many institutions at home and abroad try to develop a variety of intelligent teaching and learning systems. The technology covers a wide range, the market is uneven, and there is a lack of unified standards and evaluation mechanism (Wu & Chen, 2018). They develop independently, and many data, algorithms and technologies are exclusive to the organization, resulting in Matthew effect. Theory and technology are of low level and high repetition (Chen, 2003), and almost no online platform can truly realize adaptive learning.

### **5. Discussion and Conclusions**

#### **5.1. Discussion**

In view of the outstanding problems of intelligence, adaptability and privacy in adaptive learning research, we propose the following strategies from the aspects of cognition, technology and education combination, emotional breakthrough, learning data management and privacy protection.

##### **5.1.1. Cognitive breakthrough**

###### **(1) Cognitive breakthroughs in the brain, combining technology and brain science**

In order to achieve intelligent, adaptive and personalized functional requirements, the adaptive learning system needs to understand the learning process of human beings and the thinking process of human brains. The first step is to understand the mathematical principles and computational models of brain information processing, that is, to build computational models that can perform cognitive tasks and explain brain information processing. Therefore, it is necessary to deeply understand the mechanism of the input, transmission, exchange and output of human brain information, that is, how to produce various brain cognitive functions such as sensation and perception, emotion, choice and language. In addition, it is important for us to fully understand the information cognition and processing process of human brain. At present, people's understanding of the brain is still very limited, so it is necessary to rapidly develop the synchronization technology of information acquisition between whole brain cognition and local response (AHA!, 2020). Only by fully recognizing the brain can we achieve the breakthrough of adaptive learning technology.

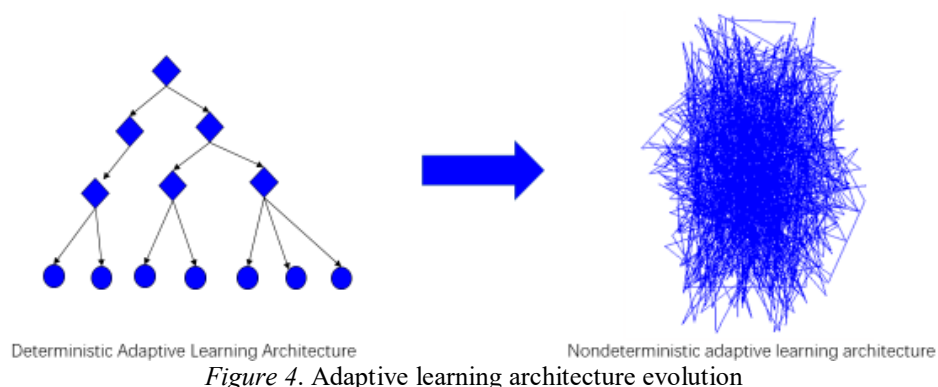
###### **(2) Cognitive breakthroughs in technology, developing zones of proximal development of technology**

On one hand, we should be clearly aware that technology is not everything. A man does what a man should do, and a machine does what a machine should do. Therefore, no matter how advanced the adaptive learning system can weaken people's thinking ability. We cannot use the intelligence of the machine to train people into a

uniform robot with the same intelligence, nor can we at the cost of human's hard work to train their own next generation into a fool at the mercy of the machine. Machines are meant to serve people, not replace them. The purpose of man-machine cooperation, in the final analysis, is to improve the quality of human life and learning effects.

On the other hand, there has always been a gap between theoretical research and practical application. According to Vygotsky's zone of proximal development theory, a reasonable technical step is set. Based on the current successful cases, adaptive learning technology needs to focus on solving two problems.

Firstly, change the deterministic learning structure to the non-deterministic learning structure, as shown in Figure 4. Traditional intelligent teaching systems and most of the current so-called "adaptive" learning systems are based on a preset learning path, and the students' learning path is almost the same. In this regard, adaptive learning should be committed to detecting students' current learning level and status through computer means, and adjusting subsequent learning contents and paths accordingly, so as to help students improve their learning efficiency. Therefore, adaptive learning realized by using artificial intelligence technology is an upgrade of traditional adaptive learning and an exploration of new learning methods. Only by changing the deterministic learning structure to the non-deterministic learning structure, different learners have different learning paths, and then personalized learning needs can be realized.



Secondly, the labeling system of knowledge points will be further improved. Only the classification of knowledge points is more detailed, the accuracy will be higher and the adaptability will be stronger. For example, the interactive teaching and learning scene of the adaptive learning calligraphy system jointly developed by our team in cooperation with Shanghai Gusuo is shown in Figure 5. The system has stored a large number of calligraphy teaching resources, including expert writing demonstration videos, famous masters' classroom videos and ancient inscriptions, including more than 60,000 inscriptions, a calligraphy library of more than 20,000 words, a calligraphy collection of more than 20,000 words, more than 10,000 videos, and 18 sets of self-developed intelligent courses. The knowledge points are subdivided into stroke, side, structure, examples and lines of writing, stroke order view, stroke position view, single hook view, double hook view, the original drawing of the tablet and other detailed content. The system can not only realize the real-time interaction between teaching and learning, but also record the learning data to the cloud synchronously. The system recommends different learning contents and paths according to students' actual level and personality differences. Students can also study independently according to their own needs. The system also sets up some experiential learning games to improve students' interest in learning and solve the problem of low efficiency in traditional calligraphy teaching and learning from multiple aspects and angles.

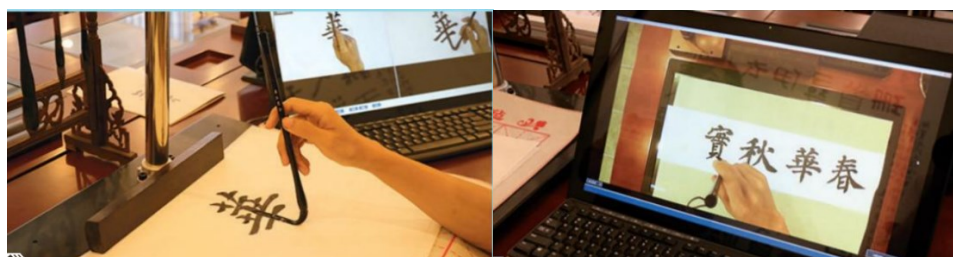


Figure 5. Adaptive learning calligraphy system teaching and learning interactive scene

### ***5.1.2. Integration of technology and education***

Adaptive learning serves learners with distinct personalities. The realization of adaptive learning needs technical support. Education has the law of education and technology has the logic of technology. However, technology is only an auxiliary means of learning, not the whole. We shouldn't rely too much on its value and effects. AI has trained Deep Blue, AlphaGo and Gaokao Robots by machine learning that surpass human beings. But we can't use machine learning to train humans. The design of any learning system should respect the essence of education, because people have the unique characteristics of human beings, including emotions, thoughts, and specific teaching rules and methods. Human characteristics and educational rules should be fully considered in the study of adaptive learning.

Educating a person is a continuous emotional process. Currently, existing AI products include Chinese tools such as photo search, hierarchical class arrangement and oral assessment, which can assist a certain learning process, but will not directly improve the quality and effect of teaching. Adaptive learning products can help to fundamentally improve the concept and way of learning only when artificial intelligence technology penetrates into each core link and the whole process of teaching.

Adaptive learning products' development needs cross-boundary collaboration and joint exploration from multiple fields and disciplines, including teaching and research experience, pedagogy, psychology, computer, big data and artificial intelligence.

### ***5.1.3. Multi-domain integration of emotional breakthrough***

A breakthrough approach is discipline integration. It is not only the integration of information technology and education, but also needs to be closely combined with brain science, psychology, statistics and other major disciplines, so that adaptive learning research can have a greater chance to make breakthroughs.

The various learning processes are interacted through polymorphic communication. The more data on the platform, the more accurate the pushing results will be. Current adaptive learning systems pay too much attention to knowledge and skills themselves, which can improve learners' speed of mastering knowledge points and test-taking ability, but it is difficult to meet human emotions and values needs. We already have millions of kinds of knowledge sample data, but the sample data on human emotions is very small, especially the data on human spirit, value and soul is almost zero (Wang, Zhao, & Wei, 2019). Therefore, strengthening the collection and sharing of the underlying data samples, especially the data of emotional value, will become the focus of the next research.

### ***5.1.4. "Code chain" management mode of learning data***

In view of the deficiency of the centralized server management mode of adaptive learning system, we first put forward the concept of "code chain" to solve the crisis of distributed processing of learning data. Code chain is the integrated innovation of graphic code technology and block chain technology. The code of "code chain" is intelligent stereo graphics code, referred to as intelligent code; and the chain of the "code chain" is equivalent to the "chain" of the traditional blockchain.

Blockchain has the advantages of distributed, decentralized, irreversible and anonymous, but it has many drawbacks, such as expansion, efficiency and security issues. The combination of smart code and blockchain is a good choice.

Intelligent code is a kind of graphic code similar to two-dimensional code, but it is better than two-dimensional code. It is essentially different from two-dimensional code. Intelligent code is to replace binary data "0" and "1" with geometry or graphics, as the text of communication between man and machine, machine and machine, in the form of three-dimensional interwoven curve geometry (graphics) for information storage, transmission and display; Geometric algorithms and structured encryption are used to manipulate storage, transmission, and interpretation of information. In the course of data collection, storage and transmission, it is convenient, decentralized, multi-dimensional and variable, personalized customization, naked eye recognition, accurate interpretation, deep encryption, intelligent anti-counterfeiting, anti-copying and traceability. The "code" in "code chain" replaces each node and block in the blockchain for distributed acquisition, storage and transmission of learning data. In addition, the code has the function of learning data security and traceability, as shown in Figure 6.

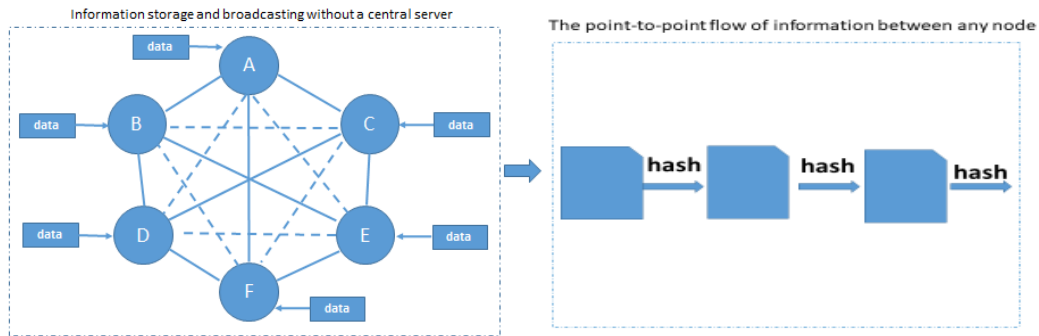


Figure 6. Decentralized data processing pattern of code chain

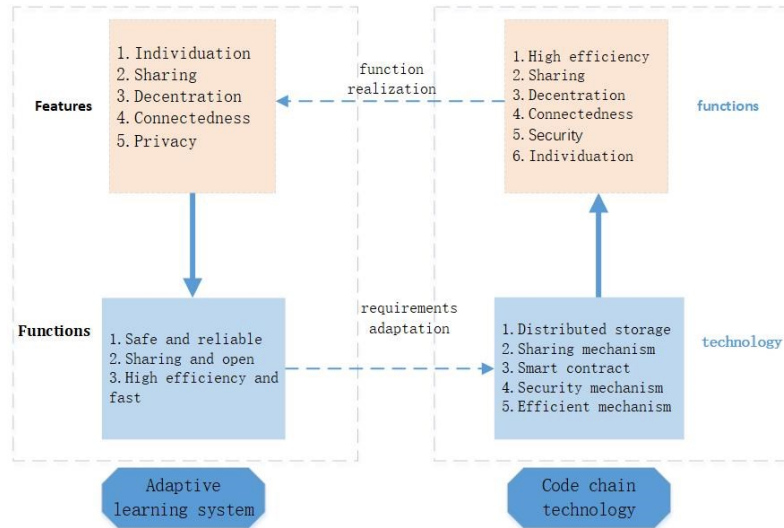


Figure 7. A way of integrating code chain technology into adaptive learning

The code chain technology compatible with the advantages of smart code and block chain is integrated into the adaptive learning system, which has the following advantages: (1) High efficiency. The acquisition of learning data is convenient and quick, the storage and transmission of learning information is larger, and the learning efficiency is improved. (2) Personalized and traceable. It can record each learner's learning process, grades and other evaluation results in the whole process. (3) Shared. Code chain technology can distribute data and computation to shareable network nodes around the world. In addition to users' private information encrypted, any learner can get the whole network data and complete backup. All participants can also get learning resources shared by different institutions or platforms, which solves the difficulty of data collection and sharing. (4) Decentralization. Without a central server, learners can establish trusted data interaction by point-to-point network communication protocol, realize distributed storage and decentralized management of learning data, and improve learning efficiency. (5) Connectivity. Each node on the code chain can maintain connectivity regardless of whether it is in the same platform and organization, which is convenient for learning data and learning resources sharing, learning content and learning peer recommendation. (6) Privacy. Each learner participates anonymously, without the need for public identity, which reduces the risk of disclosure of learners' personal information (Li, 2019).

The reason why code chain can serve adaptive learning system is that code chain technology can provide key support for adaptive learning data management and meet the functional requirements of efficient and safe learning of adaptive learning system. The way of integrating code chain into adaptive learning is shown in Figure 7.

It can be seen that with the help of code chain technology, the adaptive learning system can change from centralized to distributed, and can better realize the sharing, circulation and management of learning data in a decentralized environment.

### 5.1.5. A new adaptive learning pulling model

The principle of artificial intelligence adaptive learning builds learning models and outputs learning suggestions on the basis of Big data. At the present stage, “collecting big data -- building learning models -- outputting learning suggestions” is basic procedure to realize adaptive learning of artificial intelligence. The construction process of learning model is very complex. First of all, it needs to find out learning rules from a large amount of learning data and infinite function nesting relations. Secondly, the model is constantly trained and optimized. Finally, the study recommendation and prediction are made. The more time students spend in using the system, the more behavioral data they leave behind, and the more efficient the system becomes.

One of the core technologies of adaptive learning is learning recommendation technology. Through machine learning method, learners’ personal information, learning process data, learning style and emotional state are analyzed, and then suitable learning content or learning plan is recommended to learners.

#### (1) The traditional recommendation model of adaptive learning

At present, the learning analysis of the adaptive learning system is all conducted on the server side. This process inevitably involves the privacy of learners, the security of learning data and the right to use, as shown in Figure 8. To solve this problem, we design a new type of machine learning security model.



Figure 8. The traditional recommendation model of adaptive learning

#### (2) A new adaptive learning pulling model

Different from the traditional adaptive learning recommendation model, we design a new adaptive learning pulling model using machine learning and data pulling technology, as shown in Figure 9. “Adaptive learning pulling” contains two connotations. One is that it is different from the current use of adaptive learning recommendation system on the server side of the implementation of learning analysis, learning analysis of the adaptive learning pulling model can be implemented in the client, i.e., the learner’s computer or handheld mobile terminal. The other is that learners have a certain ability to select the recommended information, and can remotely “pull” the information recommended by the server, which is the learning content they are really interested in or need. It is divided into the following five steps:

Step 1. The learning data does not need to be uploaded to the Internet. The AI data analyzer realizes personalized analysis of learners on the student side.

Step 2. The AI data analyzer will transmit the analyzed data to the filter and display device. Learners can filter and edit the menu of learning needs according to their own actual learning needs. Unneeded learning items can be deleted, and then the system will automatically submit them to the intelligent assistant.

Step 3. The intelligent assistant blindly processes the personal information of learners and upload only learning needs to the server of the adaptive learning system as an agent.

Step 4. The adaptive learning server calculates according to the needs of the intelligent agent, and then transfers the relevant learning content or learning scheme to the intelligent agent to complete the learning pulling. At this stage, the server does not need to know who the learner is.

Step 5. The intelligent agent transmits the pulling results to the filter and display, and present them to the learner.

The “decentralized” adaptive learning pulling model has obvious advantages over the traditional learning recommendation model. First, learning analysis takes place on a student side, reducing the risk of privacy breach. Secondly, the learning analysis is not focused on the adaptive learning server, but distributed on each student side, which reduces the pressure on the server and improves the efficiency. Moreover, it selectively recommended the necessary learning content and screened out the unnecessary junk information.



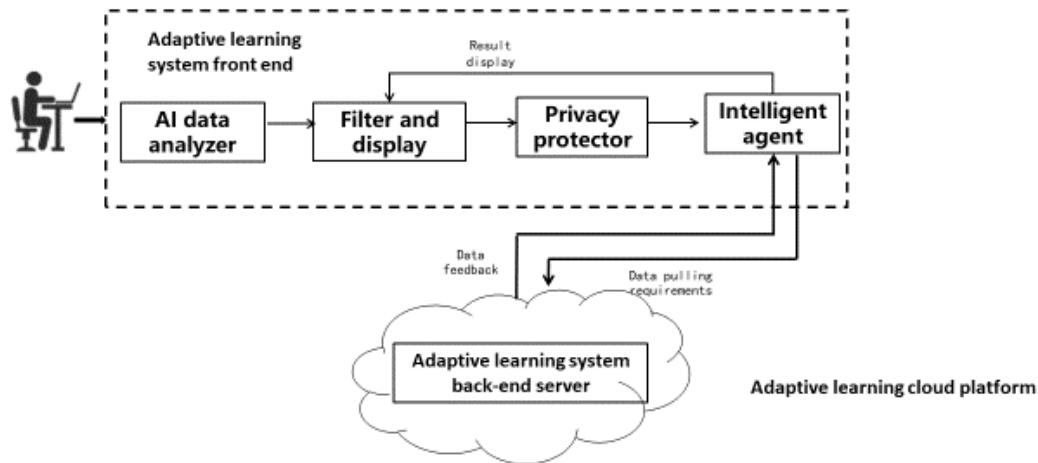


Figure 9. A new adaptive learning pulling model

## 5.2. Conclusions

Adaptive learning helps to realize personalized and autonomous learning, and is an important development direction of both online education and learning science. For more than half a century, scholars at home and abroad have done conducted extensive research on intelligent teaching and adaptive learning, and acquired a large number of achievement. However, there are still many shortcomings. Based on the systematic analysis of the connotation and progress of adaptive learning, this paper gives a new definition of adaptive learning. By literature analysis, it also discusses the major challenges of adaptive learning research, such as the lack of both knowledge of brain and technology, the bottleneck of emotion domain model, the separation of education and technology, the security and privacy risk of data management. The above challenges can be summarized into two aspects: one is mechanical, and the other is privacy. Different from the traditional perspective of adaptive learning, we open a new window and put forward some countermeasures from the perspectives of cognitive principle, learning data management and learning data security. In view of the current adaptive learning centralized management, the concept of code chain and decentralized management mode based on code chain are proposed. Different from the traditional learning recommendation technology, a new learning data pulling model based on privacy protection is proposed.

Due to the current cognitive limitations of artificial intelligence and brain learning, the research of adaptive learning is in the primary stage of development, waiting for the iterative update of theory and technology. In order to realize individualization, intelligence, autonomy, adaptability and security, it needs continuous attention of researchers, cross-integration of multi-disciplines and multi-fields, collection and sharing of massive data, and consideration of personal privacy.

## Acknowledgements

This study was funded by the National Social Science of China. The project ID was 16BGL003. It was also funded by the Social Science of the Ministry of Education of China under contract number 14YJA880033.

## References

- AHA! (2020). *Adaptive hypermedia for all*. Retrieved from <http://aha.win.tue.nl/>
- Alrifai, M., Gennari, R., & Vittorini, P. (2012). Adapting with evidence: The Adaptive model and the stimulation plan of TERENCE. *Nature Genetics*, 8(8), 328–32.
- Ausubel, D. P. (1960). The Use of advance organizers in the learning and retention of meaningful verbal material. *Journal of Educational Psychology*, 51, 267-272.
- Brusilovsky, P. (1992). Intelligent tutor, environment and manual for introductory programming. *Educational Technology & Training*, 29(1), 26–34.

- Brusilovsky, P. (1996). Methods and techniques of adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 6(2-3), 87–129.
- Brusilovsky, P., & Rizzo, R. (2003). Accessing web educational resources from mobile wireless devices: The Knowledge sea approach. *Access WS 2003, LNCS 2954*, 54–66.
- Brusilovsky, P., Eklund, J., & Schwarz, E. (1998). Web-based education for all: A Tool for developing adaptive courseware. *Computer Networks and ISDN Systems*, 30(1-7), 291–300.
- Brusilovsky, P. (2012). *Adaptive hypermedia for education and training*. Cambridge, United Kingdom: Cambridge University Press.
- Chandrasekaran, B., Johnson, T. R., & Smith, J. W. (1992). Tak-structure analysis for knowledge modeling. *Communications of the ACM*, 35(9), 124–137.
- Chang, Y. C., Kao, W. Y., Chu, C. P., & Chiu, C. H. (2009). A Learning style classification mechanism for learning. *Computers & Education*, 53(2), 273–285.
- Chen, P. D. (2003). *Research on web-based adaptive learning support system* (Unpublished doctoral dissertation). Guangzhou, China: South China Normal University.
- Chen, X., Xie, H., & Hwang, G. J. (2020). A Multi-perspective study on artificial intelligence in education: grants, conferences, journals, software tools, institutions, and researchers. *Computers and Education: Artificial Intelligence*, 1. doi:10.1016/j.caeai.2020.100005
- Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 1. doi:10.1016/j.caeai.2020.100002
- Chrys Af Iadi, K. & Virvou, M. (2012). Evaluating the integration of fuzzy logic into the student model of a web-based learning environment. *Expert systems with applications*, 39(18), 13127–13134.
- 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.
- Cui, X. P., & Xu, J. (2019). Application, issues, and trends of adaptive learning technique: An Interview with Professor David Stein, Ohio State University. *Open Education Research*, 25(5), 4-10.
- Dawson, S., Heathcote, L., & Poole, G. (2010). Harnessing ICT potential: The adoption and analysis of ICT systems for enhancing the student learning experience. *International Journal of Educational Management*, 24(2), 116-128.
- Durlach, P. J., & Lesgold, A. M. (2012). *Adaptive technologies for training and education*. Cambridge, United Kingdom: Cambridge University Press.
- Farzan, R., & Brusilovsky, P. (2008). AnnotatEd: A Social navigation and annotation service for web-based educational resources. *New Review in Hypermedia and Multimedia*, 14(1), 3-32.
- Frauke, V., Segers, E., Takashima, A., & Verhoeven, L. (2017). Effects of a tablet game intervention on simple addition and subtraction fluency in first graders. *Computers in Human Behavior*, 72(7), 200-207.
- Hartley, J. R. & Sleeman, D. H. (1973). Towards more intelligent teaching systems. *International Journal of Man-Machine Studies*, 5, 215–236. doi:10.1016/S0020-7373(73)80033-1
- Hu, W., & Chen, Y. (2018). 自适应学习:大数据时代个性化学习的新推力[Adaptive learning: A New thrust of personalized learning in the era of big data]. *Information Technology in Education in China*, 21, 42-47.
- Huang, G., Liu, X. Z., & Xue, H. Q. (2020). Intelligent and trustworthy data service system for Social-Cyber-Physical convergence. *Communications of CCF*, 16(4), 10-14.
- Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Computers & Education: Artificial Intelligence*, 1. doi:10.1016/j.caeai.2020.100001
- iResearch (2018). 2018 年中国人工智能自适应教育行业研究报告 [2018 China AI adaptive education industry research report]. Retrieved from [http://report.iresearch.cn/report\\_pdf.aspx?id=3167](http://report.iresearch.cn/report_pdf.aspx?id=3167)
- Jiang, Q., Zhao, W. & Wang, X. D. (2011). Design of ontology reference specification for user model and knowledge model in adaptive learning system. *Modern Distance Education*, 1, 61-65.
- Lan, L. N. (2015). 新加坡教育信息化现状梳理与分析 [Combing and analysis of the status quo of educational informatization in Singapore]. *The Chinese Journal of ICT in Education*, 4, 36–41.
- Li, F. Y. (2019). Blockchains-embedded learning cyber space: integrated means, functions and management patterns. *Distance Education Journal*, 37(6), 72-80.
- Li, M. (2008). Vision for education in the UK by 2020. *Social Work*, 1, 55–55.

- Li, J. Z., Dong, T. S., & Tang, J. (2020). Knowledge representation and reasoning based on spatial Cognition. *Communications of CCF*, 16(8), 18–22.
- Li, Z., Dong, X., Zhou, D., & Tong, T., (2019). A Collaborative approach for building knowledge graph in adaptive learning systems. *Modern Educational Technology*, 29(10), 80-86.
- Liu, B. (2016). A Review on the development of adaptive learning systems. *Educational Information Technology*, 9, 78–80.
- Liu, F. J. (2011). Design strategy of domain model in self-adaptive learning system of colleges and universities. *China Adult Education*, 11, 114-115.
- Nichols M. (2020). Transactional distance and adaptive learning: Planning for the future of higher education. *Open Praxis*, 12(1), 155–157.
- Papanikolaou, K. A., Grigoriadou, M., Kornilakis, H., & Magoulas, G. D. (2003). Personalizing the interaction in a web-based educational hypermedia system: The Case of inspire. *User-Modeling and User-Adapted Interaction*, 13(3), 213–267.
- Piao, Z. H. (2012). On education reform: A Study of South Korea's smart education strategy. *Education Science*, 28(4), 87–91.
- Qiu, B. S., Zhao, W., & Liu, X. Q. (2008). The Research of user model in adaptive learning system based-on semantic web. *Open Education Research*, 4, 106-111.
- Romero, V., & Zafra, E. (2009). Applying web usage mining for personalizing hyperlinks in web-based adaptive educational systems. *Computers & Education*, 53(3), 828–840.
- Somers, J. (2017). Is AI riding a one-trick pony? *Technology Review*, 120(6), 29-36.
- Surge, E. (2016). *Decoding adaptive*. London, UK: Pearson.
- Tang, J. (2020). Cognitive graph: The Next treasure of AI. *Communications of CCF*, 16(8), 8–10.
- The Ministry of Education of China. (2012). 教育信息化十年发展规划(2011-2020 年) [Ten-year development plan for educational informatization (2011-2020)]. Retrieved from [http://old.moe.gov.cn/publicfiles/business/htmlfiles/moe/s5892/201203/xxgk\\_133322.html](http://old.moe.gov.cn/publicfiles/business/htmlfiles/moe/s5892/201203/xxgk_133322.html)
- Vesin, B., Ivanovi, M., Klanja-Milievi, A., & Budimac, Z. (2012). Protus 2.0: Ontology-based semantic recommendation in programming tutoring system. *Expert Systems with Applications*, 39(15), 12229-12246.
- Walkington, C. A. (2013). Using adaptive learning technologies to personalize instruction to student interests: the impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology*, 105(4), 932-945.
- Wang, C., & Wang, D. (2014). 大学英语自适应学习环境下的学习者学习风格研究 [A Study on learners' learning styles of college English in adaptive learning systems environment]. *China Educational Technology*, 7, 145–149.
- Wang, L. P., Zhao, W., & Wei, J. H. (2019). 自适应学习系统中开放性学习者模型实证研究 [Empirical research on open learner model in adaptive learning system]. *Journal of Jilin University (Information Science Edition)*, 37(5), 512–517.
- Wang, Y. G., Xu, J. Q., & Ding, J. H. (2020). 教育 4.0 全球框架：未来学校教育与模式转变——世界经济论坛《未来学校：为第四次工业革命定义新的教育模式》之报告解读 [The Global framework of education 4.0: Future school education and mode transformation: An Analysis of the world economic forum report of schools of the future: Defining new models of education for the fourth industrial revolution]. *Distance Education Journal*, 3, 3–14.
- Weber, G., & Brusilovsky, P. (2001). ELM-ART: An Adaptive versatile system for web-based instruction. *International Journal of Artificial Intelligence in Education*, 12, 351–354.
- Wolf, C. (2003). Towards “learning style” based e-learning in computer science education. In *Proceedings of the Australasian Computing Education Conference* (pp. 273–279).
- World Economic Forum. (2020). *Schools of the future: Defining new models of education for the fourth industrial revolution*. Retrieved from <https://www.weforum.org/reports/schools-of-the-future-defining-new-models-of-education-for-the-fourth-industrial-revolution>
- Wu, W. M. & Chen, J. Y. (2018). 国内自适应学习系统的研究现状综述 [Review of research status of adaptive learning system in China]. *Jiangsu Commercial Forum*, 3, 120–124.
- Xie, H., Chu, H. C., Hwang, G. J., & Wang, C. C. (2019). Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017. *Computers & Education*, 140(10), 1-16.
- Wu, Z. H. (2020). Cybrain: Building superbrain for humans. *Journal of Zhejiang University (Engineering Science)*, 54(3), 425-426.

- Wu, Z. H., & Pan, G. (2020). 类脑研究：概念、内容及挑战 [Brain-like research: Concepts, content, and challenges]. *Communications of CCF*, 16(4), 43-48.
- Xu, K. (2020). 自适应学习的实践与探索 [The practice and exploration of adaptive learning]. *Communications of CCF*, 16(3), 58-61.
- Xu, P., & Wang, Y. N. (2011). Research status and reflection on self-adaptive learning system in China. *Modern Distance Education*, 133 (1), 25-27.
- Xu, P., Wang, Y. N., Liu, Y. H., & Zhang, H. (2013). The Learning innovation from the perspective of big data: An analysis of the U.S. report of enhancing teaching and learning through educational data mining and learning analytics and its enlightenment. *Distance Education Journal*, 6, 11-17.
- Zhao, X. K., Xu, X. D., & Long, S. R. (2015). B/S 模式下自适应学习系统个性化推荐服务研究 [An empirical study of personalized recommendation by adaptive learning system in the environment of B/S model]. *Distance Education in China*, 10, 71-78+80.
- Zhu, X. M. (1997). 人类的自适应学习-示例学习的理论与实践 [Human adaptive learning – The theory and practice of example learning]. *Open University of China Press*, 122-130.
- Zou, D., & Xie, H. (2018). Personalized word- learning based on technique feature analysis and learning analytics. *Educational Technology & Society*, 21(2), 233-244.

# A Bayesian Classification Network-based Learning Status Management System in an Intelligent Classroom

Chuang-Kai Chiu<sup>1</sup> and Judy C. R. Tseng<sup>2\*</sup>

<sup>1</sup>College of Teacher Education, Wenzhou University, China // <sup>2</sup>Department of Computer Science and Information Engineering, Chung Hua University, Taiwan // 20170006@wzu.edu.cn // judycrt@chu.edu.tw

\*Corresponding author

**ABSTRACT:** Awareness of students' learning status, and maintaining students' focus and attention during class are important issues in classroom management. Several observation instruments have been designed for human observers to document students' engagement in class, but the processes are time-consuming and laborious. Recently, with the development of artificial intelligent technologies, artificial intelligence in education (AIED) has become an important research topic. Several studies have applied image recognition technologies to determine students' learning status. However, little research has employed both sensor technology and image recognition technology in learning status analysis. Moreover, it remains unknown if learning status analysis is accurate enough to substitute for human observers. Furthermore, no feedback has been provided individually to students to manage their learning status by maintaining their attention in class. In this paper, a learning status management system in an intelligent classroom is proposed. Several types of information about students were detected and collected by both sensor technology and image recognition technology, and a Bayesian classification network was employed to inference the students' learning status. Moreover, the system includes a feedback mechanism, which not only provides the results of the just-in-time learning status analysis to teachers, but also notifies students who are detected as being unfocused in class. Two experiments were conducted to verify the accuracy and effectiveness of the proposed system. Results showed that the learning status analysis highly corresponded to the observation of human beings, and the students were more attentive in class.

**Keywords:** Classroom management, Intelligent classroom, Learning status analysis, Bayesian classification network

## 1. Introduction

In traditional classrooms, learning efficiency is usually influenced by students' learning status. If students are inattentive, drowsy, or even fall asleep, they are not able to absorb the content taught by teachers. Teachers usually use a wide variety of classroom management strategies to keep students focused and attentive during class (Kounin, 1970; Evertson, 1994; Kyriacou, 1997). However, since teachers must pay attention to their own instruction, it is challenging for them to also be aware of the individual learning status of each student (Yang, Cheng, & Shih, 2011) and to provide suitable feedback in a timely manner. It is also impossible for teachers to record students' individual learning status all the time in-class for further evaluation and/or analysis. While several classroom observation instruments have been designed for human observers to document students' engagement in class (O'Malley et al., 2003; Dockrell, Bakopoulou, Law, Spencer, & Lindsay, 2012; Eddy, Converse, & Wenderoth, 2015), the observation and documentation processes mainly depend on human labor. It is not only time-consuming, but also laborious. Moreover, since the learning status is recognized by observers rather than teachers, teachers are not able to learn the just-in-time results of the observation and change their instructional strategies accordingly to achieve better classroom management.

Recently, with the development of artificial intelligent (AI) technologies, artificial intelligence in education (AIED) has become an important research topic (Hwang, Xie, Wah, & Gašević, 2020; Chen, Xie, Zou, & Hwang, 2020; Chen, Xie, & Hwang, 2020; Tang, Chang, & Hwang, 2021; Yang, Ogata, Matsui, & Chen, 2021). Chen et al. (2020) attempted to investigate the gap between application and theory during the rise of AIED; one of their findings was that “*most influential AIED studies are concerned about the application of AI technologies in the contexts of online or web learning, while few concerned about the promotion of learning and teaching in physical contexts with the help of AI technologies*” (p. 16). Their finding reveals that applying AI technologies in physical classroom settings for enhancing the learning and teaching process is a potential research issue. In view of this, research on intelligent classrooms which employ AI technologies, such as sensor technology and image recognition technology, has arisen (Zhu, Xu, & Gao, 2020; Li, Tan, & Hu, 2021; Li, 2021). Generally, the term “Intelligent classroom” refers to a physical classroom that integrates advanced educational technology to improve teachers' abilities to promote student learning and students' abilities (Winer & Cooperstock, 2002; Ramadan, Hagrass, Nawito, El Faham, & Eldesouky, 2010).

To address the problem of learning status management in class, some research has employed image recognition technologies to analyze the videos/images of students, using facial actions and expressions to determine students' learning status in real time (Hwang & Yang, 2009; Yang, Cheng, & Shih, 2011; Huang, Li, Qiu, Jiang, Wu, & Liu, 2020; Yang, Yao, Lu, Zhou, & Xu, 2020). However, students' learning status is not only reflected in their facial actions and expressions. Although sensor technology is useful for detecting students' behaviors in the classroom (Chang & Chen, 2010), little research has employed sensor technology in learning status analysis. Moreover, most studies did not evaluate the accuracy of the learning status analysis by comparing it with judgements by classroom observers. It is therefore uncertain whether the results of learning status analysis are sufficiently accurate to substitute for human observers. On the other hand, to keep students attentive in class, feedback should be provided to both students and teachers according to the learning status detected. However, only some research has provided feedback to teachers, while little research has provided feedback individually to the students themselves to maintain their attention in class.

To create an intelligent classroom with a more effective classroom management facility, a learning status management system is proposed in this study. Various types of sensors were used to obtain students' physiological signals, and a small camera was installed in front of each desk to capture the image or take videos of each student. Several features that could be used to infer students' learning status were detected and collected by sensor technology and image recognition technology. To infer students' learning status from the collected features, a Bayesian classification network was employed. A Bayesian classification network is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG) (Jensen, 1996). It is ideal and versatile for a wide range of tasks including prediction, diagnostics, reasoning, and decision making in situations of uncertainty (Pourret, Naïm, & Marcot, 2008). The learning status of students inferred by the proposed system could be recorded for further analysis. Moreover, a feedback mechanism was also included in the system to notify students who had become inattentive, drowsy or had fallen asleep so as to regain their attention. It also provided a dashboard for teachers to visualize the real-time learning status of each student; teachers could then adjust their instructional strategies in a timely fashion so as to achieve better classroom management.

To evaluate the performance of the proposed system in classroom management, system validation was performed to verify the accuracy of the learning status management system. The correlation between the students' learning status determined by the proposed system and that determined by human observers was analyzed in the system validation. Moreover, a quasi-experiment was conducted to evaluate the efficacy of the learning status management system. Two classes of students taking the course "Introduction to Computer Science" participated in the experiment. One class was assigned to the experimental group, which studied in the intelligent classroom with the learning status management system enabled. Another class was assigned to the control group, which also studied in the same classroom with the learning status management system disabled. The degrees of students' attention of the two classes were analyzed and compared. Thus, there were two research questions to be investigated in this study:

Q1. Does the learning status determined by the proposed system correspond to that determined by human observers?

Q2. Can students' attention in class be promoted when the proposed system is enabled?

## 2. Literature review

### 2.1. Classroom management

Classroom management, also known as class management, covers a very wide range of activities (Evertson, 1994). Doyle (1986) defined classroom management as the necessary preparation and procedures for establishing and maintaining an environment in which teaching and learning take place. He believed that classroom management is a prerequisite for successful teaching. Froyen (1988) defined classroom management as including content management, covenant management and conduct management. Content management refers to the management of classroom space, teaching materials, equipment, the movement of students, and the process of instruction. Covenant management focuses on the classroom group as a social system; teachers should pay attention to managing interpersonal relationships in the classroom. Conduct management refers to dealing with discipline problems in the classroom. Emmer and Stough (2001) defined classroom management as "*actions taken by the teacher to establish order, engage students, or elicit their cooperation*" (p. 103) The Glossary of Education Reform provided a versatile concrete definition of classroom management as "the wide variety of skills and techniques that teachers and schools use to keep students organized, orderly, focused, attentive, on task, and academically productive during a class" (Great Schools Partnership, 2014)

Evertson and Weinstein (2006) believed that in order to attain high quality classroom management, five actions are indispensable for teachers: (1) establish a caring and supportive relationship with students; (2) organize and implement teaching to optimize students' learning opportunities; (3) encourage students to participate in academic tasks; (4) promote students' social skills and self-regulation ability; and (5) use appropriate interventions to help students solve their behavior problems. Kyriacou (1997) identified that the most common and destructive problem behaviors were talking with classmates, followed by inattention, wandering, and idleness. The findings indicated that relatively minor forms of student misbehaviors are a common concern for teachers, and that teachers spend a considerable amount of time on behavior management issues (Clunies-Ross, Little, & Kienhuis, 2008).

From the literature above, it can be seen that how to improve the effectiveness and efficiency of classroom management, which involves identifying students' learning behaviors to determine their learning status and taking suitable actions to help them concentrate on learning, has become an important research topic. In this study, the term *learning behavior* refers to students' behaviors that occur during the learning process. The term *learning status* refers to an individual's mental state during the learning process, which can be determined by the individual's learning behaviors. For example, a student with the learning behavior of "talking with classmates" while the teacher lectures would be considered as having the learning status of "inattention."

## **2.2. Learning behavior identification to assist classroom management**

Delgado et al. (2011) indicated that concentration during learning is the key factor influencing learning effect. If a student cannot concentrate on learning, it will affect the learning mood, resulting in lower learning concentration and lower learning effect. Schmidt (1990) also pointed out that attention plays an important role in traditional classroom learning. When students start to lose concentration or feel tired or even start to fall asleep, the learning content will be ignored and the learning efficacy will be decreased. To help students concentrate on learning, teachers should pay attention to classroom management, especially to the management of students' learning status, which can be determined by identifying their external learning behaviors.

To identify and record students' learning behaviors in a physical classroom for learning status analysis, several tools have been developed in the literature, such as classroom observation instruments, classroom teaching video analysis software, and/or observation scales (O'Malley et al., 2003; Dockrell, Bakopoulou, Law, Spencer, & Lindsay, 2012; Eddy, Converse, & Wenderoth, 2015; Flanders, 1961; Rich & Hannafin, 2009). For instance, the Flanders Interaction Analysis System (FIAS) is an observational tool used to observe verbal communication in the classroom (Flanders, 1961). It uses a system of categories to encode the classroom behavior of both teacher and students. However, non-verbal gestures are not taken into account (Amatari, 2015). Classroom Video Analysis (CVA) is another well-known method in which the entire teaching process is recorded and then analyzed (Kersting, 2008; Kersting et al., 2012). CVA measures "usable teacher knowledge" by scoring their written analyses of classroom video clips.

These traditional methods of learning behavior identification for learning status analysis rely heavily on the manpower of the classroom observers, so the process is rather time-consuming, laborious and inefficient. Moreover, since the analytical results cannot be provided to the teachers in a timely manner while they are instructing students in the classroom, they are not able to adjust their instruction strategies immediately to achieve better classroom management.

## **2.3. AI and Sensor technology for learning status analysis**

With the development of Artificial Intelligence (AI), various AI technologies, such as sensor technology, image recognition technology, Bayesian classification networks, fuzzy logic, decision trees, neural networks, genetic algorithms, and Hidden Markov Models (HMM), have been employed in the education domain (Tang, Chang, & Hwang, 2021). To eliminate the timely constraint and relieve the burden of manpower in traditional learning status analysis, some studies have applied AI technologies to develop systems for learning status analysis (Hwang & Yang, 2009; Yang, Cheng, & Shih, 2011; Huang, Li, Qiu, Jiang, Wu, & Liu, 2020; Yang, Yao, Lu, Zhou, & Xu, 2020).

Hwang and Yang (2009) proposed an auto-detection and reinforcement mechanism for learning status analysis in distance education. They employed image recognition and detection techniques to recognize the inattention and fatigue status of learners. A Bayesian network assessment was employed in their reinforcement mechanism to reduce detection misjudgment and enhance accuracy. Yang et al. (2011) proposed a computer vision system to

automatically analyze learners' videos to recognize nonverbal facial expressions to discover the learning status of students in distance education. Adaboost classifiers were applied to extract facial parts from students' videos, and specific emotional expressions were recognized by HMM. To recognize students' typical classroom behaviors, Huang et al. (2020) applied a deep convolutional neural network (D-CNN) to analyze students' images of head poses and facial expressions. Yang et al. (2020) identified students' concentration degrees during classroom learning by detecting their head motions, such as raising and lowering their heads, from in-classroom videos. The concentration degrees are linked to the teacher's teaching characteristics, including audio features, the course topics taught in different time periods, and the speed of the teacher's speech when explaining the topics.

As we can see from the literature, most of the studies that have used AI technologies in learning status analysis employed image recognition technologies to determine students' learning status in real time. However, students' learning status can not only be reflected in their facial actions and expressions, but can also be revealed by their physiological signals, such as body movement, and pulse. Although sensor technology is useful in detecting students' behaviors in the classroom (Chang & Chen, 2010), little research has employed sensor technology in learning status analysis. Moreover, most studies did not evaluate the accuracy of the learning status analysis. While some studies have evaluated the accuracy, the evaluations were only based on testing examples of facial recognition. No comparison with human judgements using real images captured in the physical classroom has been made. It still remains unknown if learning status analysis is sufficiently accurate to substitute for human observers. On the other hand, to manage students' learning status to maintain their attention in class, feedback should be provided to both students and teachers according to the learning status detected. However, only some research has provided feedback to teachers to allow them to consider changing their instructional strategies. Little research has provided feedback individually to students to manage their learning status, keeping them attentive in class.

To fill the research gap, a learning status management system is proposed in this paper. Both sensor technology and image recognition technology are employed for learning status analysis. To validate the accuracy of learning status analysis, the correlation between the students' learning status determined by the proposed system and those determined by human observers was analyzed. A feedback mechanism, which will provide feedback to both the teachers and the students, is also included to keep the students attentive. With the help of the proposed system, it is hoped that better classroom management can be achieved.

### **3. Method**

#### **3.1. Bayesian classification network-based learning status management system**

The proposed learning status management system included a learning status inference engine and a feedback mechanism. The learning status inference engine was responsible for analyzing students' learning status. The determined learning status was recorded in a database. The feedback mechanism was responsible for giving suitable feedback to both teachers and students according to the students' learning status. When students received feedback, they would be aware of their learning status and adjust it so as to be attentive. When teachers received feedback, they could change their instruction strategies to maintain students' attentiveness.

A four-layer Bayesian inference network is employed in the learning status inference engine. A Bayesian network is a type of probabilistic graphical model that uses Bayesian inferencing for probability computations. A set of variables and their conditional dependencies are represented via a directed acyclic graph in the Bayesian network. Bayesian network assessment can reduce detection misjudgment and enhance accuracy. It was found that Bayesian networks could also be used to evaluate or predict the learning behavior of students in a distance learning environment (Xenos, 2004; Hwang & Yang, 2009).

As shown in Figure 1, the four-layer Bayesian classification network is composed of a sensor layer, a feature layer, a behavior layer and a status layer. The sensor layer consists of several types of sensing devices, such as microphone, camera, body temperature, and so on. The features of a learner can be captured and recognized via these sensors. Differing from past studies, the Bayesian classification network proposed here not only uses image recognition technology to incorporate the features that can be recognized from the images/video captured by camera, but also considers the information captured from sensors embedded in the classroom and worn by students. According to the features obtained, the students' behaviors are inferred and determined. Misbehavior refers to the behaviors that would distract other students from their learning, such as chatting with classmates, bad posture or leaving their seats. For instance, the frequencies of a learner's eyes being half-closed and head



nodding can be obtained by facial feature recognition from the image/video captured by camera. The drowsy behavior of a learner can then be inferred by integrating the two frequencies. If the behavior of a student is predicted as misbehavior or fatigue, the learning status of this student is recorded as inattentive, and the degree of inattention is determined by the frequency of the misbehavior or fatigue. The sensors used for detecting the conditions of students and the learning behaviors determined are listed in Table 1. The behavior layer currently includes two behaviors, misbehavior and fatigue behavior, but can be extended to meet requirements in the future.

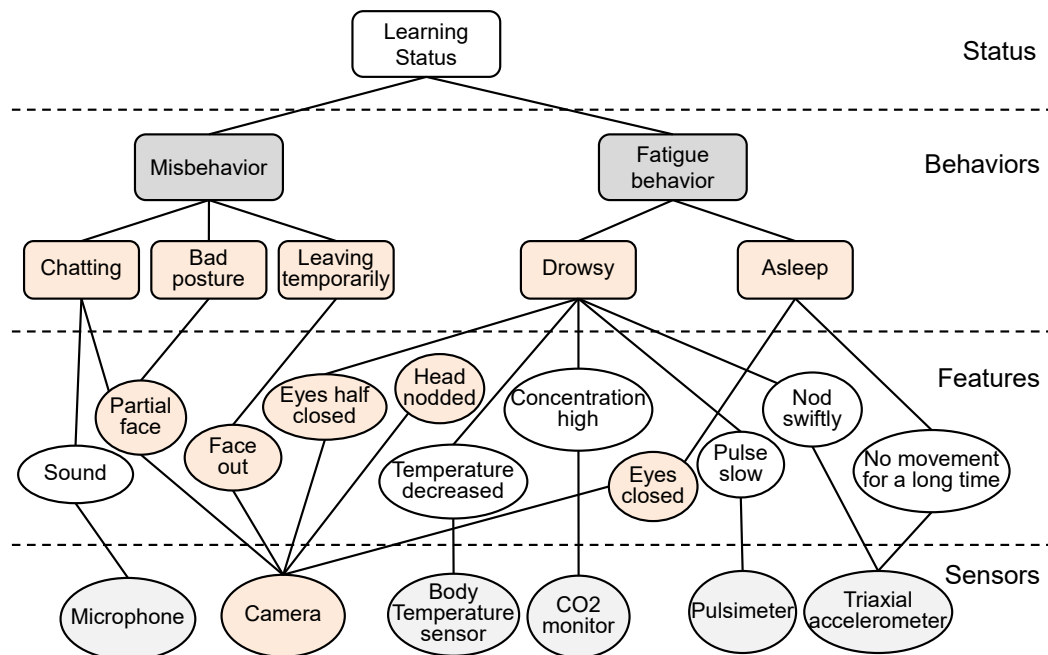


Figure 1. Bayesian classification network for learning status analysis

Table 1. The features of learning behaviors

Sensor	Condition	Behavior
Microphone	chatting	misbehavior
Camera	winking frequency and the face is not in the right place	bad posture, leaving temporarily, drowsy or asleep
Body temperature	temperature decreasing	drowsy
CO <sub>2</sub> monitor	high concentration	drowsy
Pulsimeter	pulse getting slow	drowsy
Triaxial accelerometer	head is nodding swiftly	drowsy or asleep

When a student is determined to be inattentive by the inference engine, the degree of inattentiveness is recorded in a database. The feedback mechanism gives feedback to both the teacher and the students accordingly. For inattentive students, the feedback could be a blinking LED installed in front of the student's desk, or a mild shake of the student's seat or smart bracelet, to remind him/her to be attentive. The feedback mechanism for students could be determined by the equipment installed in the intelligent classroom. In this study, we used LED lights as the feedback mechanism. For the teacher, a dashboard presenting the learning status of each student was displayed in the interface of the proposed learning status management system, as shown in Figure 2. The color of the status block for each student shows the degree of inattentiveness. A red block means very inattentive, a yellow block means inattentive, and a green block means attentive. Additionally, if the face is not detected all the time, it means the student is absent from class, and the status block is displayed as black. With the dashboard, the teacher can learn the status of all the students at a glance. If most students are inattentive, the teacher could change his/her instruction strategy to regain the students' attention. When the class is finished, lists of absent students and inattentive students are also provided. The teacher can use this information to provide special care to individual students after class.

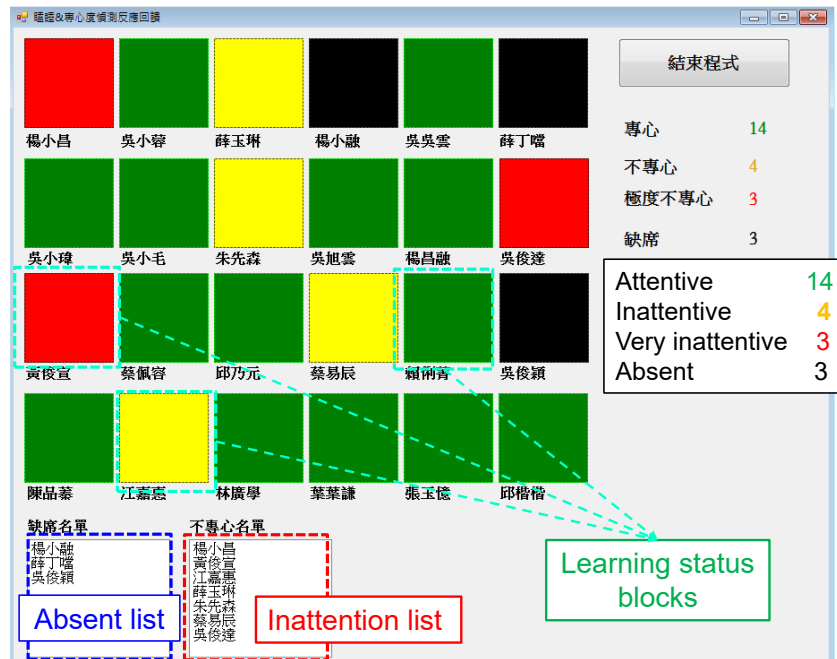


Figure 2. The interface of the learning status management system

### 3.2. Experiment design

The four-layer Bayesian inference network-based learning status management system was implemented in a context-aware classroom (Figure 3). The classroom is equipped with several sensors and feedback devices in the intelligent classroom, and Zigbee technology was employed to drive the equipment. Cameras were used to collect the features of students for learning status management, and a CO<sub>2</sub> monitor and three complex sensors were installed for collecting the context information (CO<sub>2</sub> concentration, temperature, humidity and illumination). Two experiments were conducted: one for accuracy and the other for effectiveness. The two experiments investigated two research questions, *Q1: Does the learning status determined by the proposed system correspond to that determined by human observers?* and *Q2: Can students' attention in class be promoted when the proposed system is enabled?* All participants involved in the experiment were informed in advance that their facial information would be collected and recorded during the experiment.

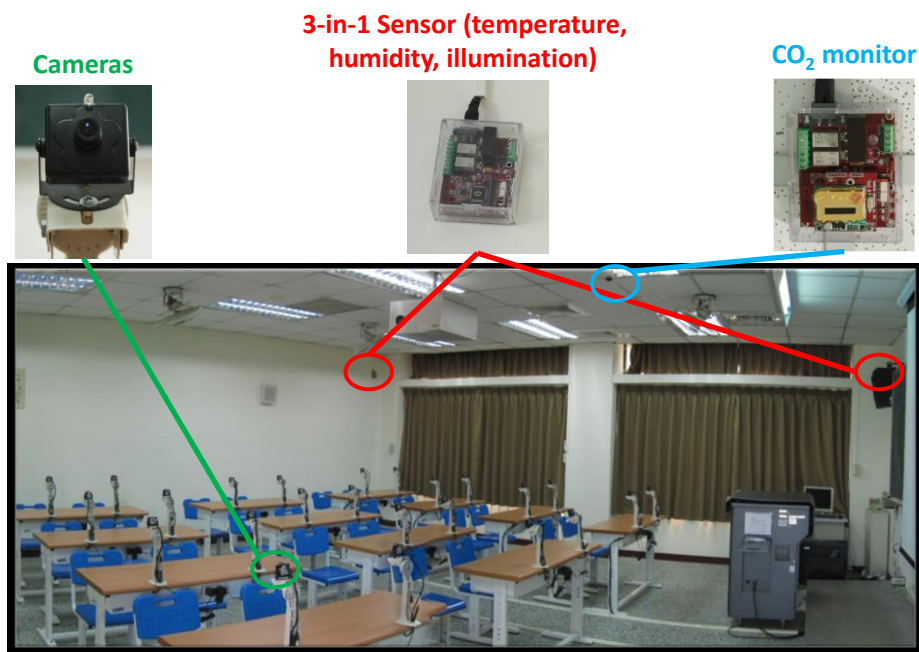


Figure 3. Intelligent classroom and embedded sensors and controllers

In order to verify the accuracy of the learning status inference engine, compared to human observers (raters), the first experiment was conducted as shown in Figure 4. There were 20 students who participated in this experiment. While they learned in the intelligent classroom with the proposed learning status management system enabled, the face of each learner was captured by the camera set before each of them. During the class, both the video clips and the learning status determined by the system were recorded. After class, each video clip of each student was manually examined by three raters, and the frequency of the fatigue state of each student was rated. The rating results were then compared with the results determined by the system.

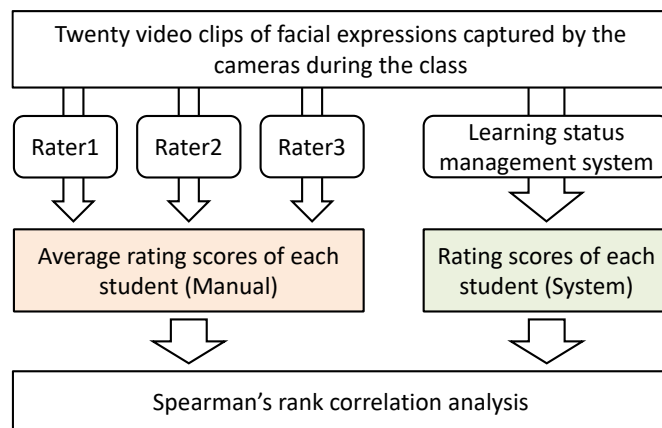


Figure 4. Accuracy evaluation procedure

On the other hand, in order to verify the effectiveness of the learning status management system, a 2-week field experiment was conducted in the intelligent classroom. Sixty-four students in two classes were involved in the experiment. The learning subject was “Introduction to computer science” and each class was 45 minutes in length. In week 1, both classes learned in the same classroom and the learning status management system was disabled during class. After class, a pre-questionnaire was administered for the students to complete. In week 2, one class was assigned to be the experimental group, and the other was assigned to be the control group. When the experimental group was learning in the classroom, the learning status management system was enabled. Conversely, the system was disabled when the control group was learning in the same classroom. Similar to the process in week 1, a post-questionnaire was administered after class for the students to complete. The experiment process is shown as Figure 5.

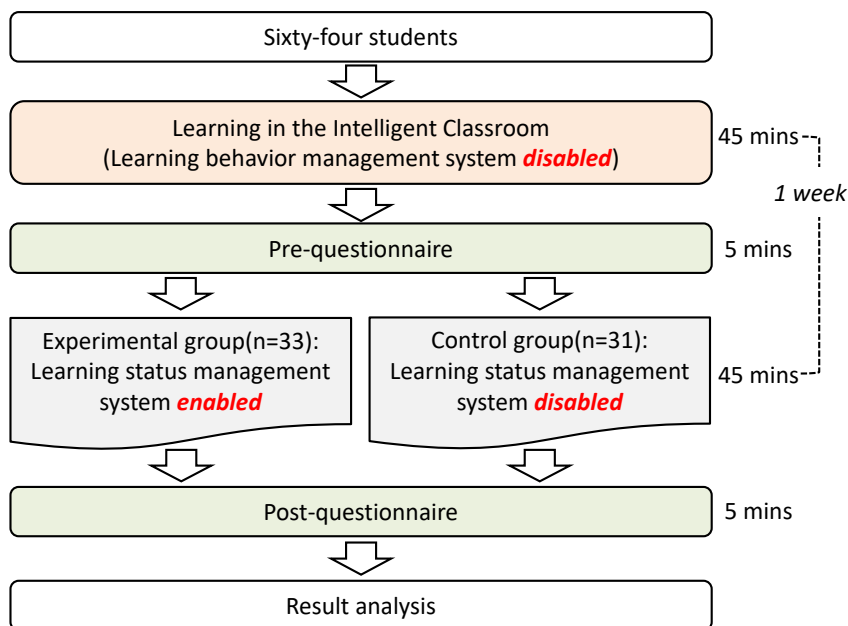


Figure 5. Effectiveness evaluation procedure

There are three question items in the pre-questionnaire and post-questionnaire for the students to self-evaluate their learning status during class, as listed in Table 2. The questionnaire used a 5-point Likert scale. Students were asked to self-evaluate their degree of conformity with “completely agree (5),” “agree (4),” “no opinion (3),”

“disagree (2)” and “completely disagree (1).” Since the purpose of this experiment was to verify whether the learning status management system could keep students attentive during learning, which is the major purpose of classroom management, the effectiveness of the proposed system was evaluated by measurements that reflected the students’ degree of attentiveness.

Table 2. Learning status questionnaire

	Question items
I1	I was mostly attentive during the class
I2	I seldom remained fatigued during the class
I3	I didn’t doze off during the class

## 4. Experiment results

### 4.1. Accuracy evaluation of the learning status inference engine

In order to evaluate the correlation between human-rated ranks and computer-rated ranks of students’ degrees of inattention, Spearman’s rank-order correlation was determined. It is a nonparametric version of the Pearson product-moment correlation. Spearman’s correlation coefficient ( $r_s$ ) measures the strength and direction of association between two ranked variables. The coefficient is computed by formula (1), where  $r_s$  represents the Spearman’s rank correlation coefficient,  $n$  represents the number of observations, and  $di$  represents the difference between the two ranks of each observation. The result of Spearman’s rank-order correlation is listed in Table 3.

$$r_s = 1 - \frac{6 \sum di^2}{n(n^2 - 1)} \quad (1)$$

Table 3. Correlation between system rating and manual rating

	System rating	Correlation Coefficient	Ranker
Spearman’s rho ( $\rho$ )		Sig.(2-tailed)	.787***
		N	.000
			20

Note. \*\*\*Correlation is significant at the 0.001 level (2-tailed).

Table 4. Explanation of the value range of the rank correlation

Range of coefficient	Correlation degree
$\rho \leq 0.3$	Low
$0.3 < \rho \leq 0.7$	Medium
$\rho > 0.7$	High

From Table 3 and Table 4, we can find that the Spearman coefficient ( $r_s$ ) is 0.787, which is larger than 0.7. The test of correlation significance shows that probability Sig. (2-tailed) is 0.000 ( $< .05$ ). This implies that there is a significant positive correlation between the system ratings and human ratings, and the correlation degree is high. From the analytical result, we can find that the rating results from the system can be treated as similar to the human rating results. In other words, the prediction of learning status by the proposed inference engine is highly accurate. We can therefore answer research question Q1: the learning status determined by the proposed system highly corresponds to that determined by human observers.

### 4.2. Effectiveness evaluation of the learning status management system

In order to investigate the effectiveness of the learning status management system, the learning status questionnaire shown in Table 2 was conducted after class in week 1 and week 2. The questionnaire results collected in week 1 were regarded as the pre-questionnaire results and those in week 2 as the post-questionnaire results. An independent sample  $t$ -test was applied to evaluate the results. The analysis results of the pre-questionnaire showed that there was no significant difference in I1,  $t(62) = -0.800$ ,  $p = .427$ ,  $d = 0.20$ , between the experimental group ( $M = 3.48$ ,  $SD = 1.00$ ) and the control group ( $M = 3.29$ ,  $SD = 0.94$ ). Moreover, there was also no significant difference in I2,  $t(62) = -0.604$ ,  $p = .548$ ,  $d = 0.15$ , between the experimental group ( $M = 3.24$ ,  $SD = 1.12$ ) and the control group ( $M = 3.06$ ,  $SD = 1.237$ ). Similarly, there was also no significant difference in I3,  $t(62) = -0.934$ ,  $p = .354$ ,  $d = 0.35$ , between the experimental group ( $M = 3.58$ ,  $SD = 1.06$ ) and the control group ( $M = 3.23$ ,  $SD = 0.96$ ).

After different treatments, the post-questionnaire was collected. The independent sample *t*-test result of the post-questionnaire between the two groups is listed in Table 5.

Table 5. Independent sample *t*-test result of the post-questionnaire between the two groups

Question items	Mean (Std.)		<i>df</i>	<i>t</i>	Effect size( <i>d</i> )
	System enabled ( <i>N</i> = 33)	System disabled ( <i>N</i> = 31)			
I1. I was mostly attentive during the class	3.70 (0.73)	3.23 (0.85)	62	-2.394*	0.59
I2. I seldom remained fatigued during the class	3.79 (0.74)	3.00 (1.32)	46.59	-2.926**	0.74
I3. I didn't doze off during the class	4.12 (0.74)	3.45 (1.18)	49.91	-2.702**	0.68
<b>Average score</b>	3.87 (0.53)	3.23 (0.96)	46.16	-3.273**	0.83

Note. \**p* < .05; \*\**p* < .01.

It was found that there was a significant difference in the average scores of the three question items,  $t(46.16) = -3.273$ ,  $p = 0.002$ ,  $d = 0.83$ , and the average score for the “System enabled group” ( $M = 3.87$ ,  $SD = 0.53$ ) was significantly greater than that for the “System disabled group” ( $M = 3.23$ ,  $SD = 0.96$ ). For I1, “I was mostly attentive during the class,”  $t(62) = -2.394$ ,  $p = .020$ ,  $d = 0.59$ , and the average score for the “System enabled group” ( $M = 3.70$ ,  $SD = 0.73$ ) was significantly greater than that for the “System disabled group” ( $M = 3.23$ ,  $SD = 0.85$ ). For I2, “I seldom remained fatigued during the class,”  $t(46.59) = -2.926$ ,  $p = .004$ ,  $d = 0.74$ , and the average score for the “System enabled group” ( $M = 3.79$ ,  $SD = 0.74$ ) was also significantly greater than that for the “System disabled group” ( $M = 3.00$ ,  $SD = 1.32$ ). Similarly, for I3, “I didn't doze off during the class,”  $t(49.91) = -2.702$ ,  $p = .009$ ,  $d = 0.68$ , and the average score for the “System enabled group” ( $M = 4.12$ ,  $SD = 0.74$ ) was significantly greater than that for the “System disabled group” ( $M = 3.45$ ,  $SD = 1.18$ ). Hence, from Table 5, we can conclude that the proposed learning status management system was able to help students be more attentive, experience less fatigue, and doze off less often during class. We can therefore answer research question Q2: students' attention in class can be promoted when the proposed system is enabled.

## 5. Discussion

In this paper, two experiments were conducted to verify the accuracy and effectiveness of the proposed system. As shown in Table 3, the result of accuracy evaluation showed that the learning status determined by the system was highly correlated with the result determined by the human observers. This finding means that the proposed system can substitute human observers, and relieve the burden of manpower in traditional learning status analysis. Moreover, since the proposed system gives feedback to both teachers and students immediately after the students' learning status is determined, the time constraint of traditional learning status analysis can be eliminated.

Besides applying AI technologies to assist teachers in recognizing students' learning status, the effectiveness of the proposed system was also evaluated. The experimental results show that the proposed learning status management system was able to help students remain attentive in class. When the proposed system was enabled, the students felt more attentive, less fatigued, and were less likely to doze off. This result could be credited to the feedback mechanism of the proposed system. Since those students who are inattentive are marked in the interface of the management system (Figure 2), teachers can easily identify the students' learning status and take action to keep students attentive. For example, when most students are inattentive, the teacher can give a quiz or tell a joke to regain their attention. If only some students are inattentive, the teacher can ask a specific inattentive student to answer a question to stimulate his/her attention. On the other hand, students who are determined to be inattentive will also receive feedback from the proposed system. That will remind them to keep attentive even when the teacher does nothing in response to their learning status.

The experimental results provide evidence of the contribution of the proposed system to classroom management, but there are nevertheless some limitations to this study. Due to the limitations of equipment, not all the sensors indicated in the proposed Bayesian classification network for learning status analysis (Figure 1) were used in the experiment. The inference power of the Bayesian classification network proposed was not fully reflected in the experimental results. Moreover, the experiment was only conducted for one week. The experimental results can only represent the students' performance in this short period of time. Furthermore, only 64 students participated in the experiment. More participants would be required to obtain stronger results.

Yang et al. (2021) indicated that smart learning environments should not only focus on performance but also human feelings. Ethics and norms should also be considered, and smart learning analytics should ensure privacy by enabling students to decide whether to give their permission for capturing and using their facial features. In this study, all participants were informed and consented that their facial information would be collected and recorded during the experiment.

## 6. Conclusion and future work

In this study, a learning status management system based on a Bayesian classification network was proposed in an intelligent classroom. Differing from past research, both sensor technology and image recognition technology were employed in the proposed system. Two experiments were conducted to evaluate the accuracy and effectiveness of the proposed system. From the experimental results, the learning status determined by the proposed system was highly correlated to that determined by human observers. Furthermore, the degrees of students' attention in class could be promoted when the proposed system was enabled. To sum up, the proposed system is helpful to teachers for ensuring more effective classroom management. As many researchers have indicated that the concentration of students' learning is the key factor influencing the learning effect (Delgado et al., 2011; Schmidt, 1990), it can be expected that with the help of the proposed system, students' learning performance will be promoted.

In the future, we will utilize all of the sensors indicated in the proposed Bayesian classification network for learning status analysis in the experiments to fully investigate the power of the proposed system. Moreover, the experiments will be conducted for at least one semester to evaluate the impacts of the system not only on learning status management but also on learning performance. Furthermore, more students will participate in the experiments to obtain stronger results.

In the post-pandemic era of Covid-19, in order to avoid face-to-face contact, many in-class learning activities have gradually transformed into online learning, either synchronous or asynchronous. How to manage students' learning status, and keep them attentive during online learning is more challenging than in classroom learning. Currently, most learning devices used in online learning are equipped with cameras and microphones. Smart bracelets that can detect various physiological signals are also becoming increasingly versatile and popular. Excluding the sensors installed in the intelligent classroom, the proposed learning status management system can also be applied in on-line learning environments. However, to reduce the communication load of transmitting the large amount of information captured by various sensors, the learning status inference engine has to be redesigned using edge computing. Moreover, the effectiveness of the proposed system needs to be further investigated in the context of on-line learning in the future.

## Acknowledgment

This study is supported in part by the Ministry of Science and Technology of the Republic of China under contract numbers MOST 108-2622-H-216-001-CC3 and MOST 109-2511-H-216-001-MY3, and by Zhejiang Provincial Natural Science Foundation of China [LY19F020037]. The authors would like to thank Dr. Cheng-Chang Lien for his help of supporting image recognition technology.

## References

- Amatari, V. O. (2015). The instructional process: A Review of Flanders' interaction analysis in a classroom setting. *International Journal of Secondary Education*, 3(3), 43-49.
- Chang, B., & Chen, C. W. (2010). Students' competitive preferences on multiuser wireless sensor classroom interactive environment. In *Proceedings of the 2010 10th IEEE International Conference on Advanced Learning Technologies* (pp. 570-572). doi:10.1109/ICALT.2010.162
- Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A Review. *IEEE Access*, 8, 75264-75278.
- Chen, X., Xie, H., & Hwang, G. J. (2020). A Multi-perspective study on artificial intelligence in education: grants, conferences, journals, software tools, institutions, and researchers. *Computers and Education: Artificial Intelligence*, 1, 100005. doi:10.1016/j.caeai.2020.100005
- Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 1, 100002. doi:10.1016/j.caeai.2020.100002

- Clunies-Ross, P., Little, E., & Kienhuis, M. (2008). Self-reported and actual use of proactive and reactive classroom management strategies and their relationship with teacher stress and student behaviour. *Educational psychology*, 28(6), 693-710.
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates.
- Delgado, M. R., Phelps, E. A., & Robbins, T. W. (Eds.). (2011). *Decision making, affect, and learning: Attention and performance XXIII*. New York, NY: Oxford University Press Inc.
- Dockrell, J. E., Bakopoulou, I., Law, J., Spencer, S., & Lindsay, G. (2012). *Developing a communication supporting classroom observation tool*. London, UK: DfE.
- Doyle, W. (1986). Classroom organization and management. *Handbook of Research on Teaching*, 3, 392-431.
- Eddy, S. L., Converse, M., & Wenderoth, M. P. (2015). PORTAAL: A Classroom observation tool assessing evidence-based teaching practices for active learning in large science, technology, engineering, and mathematics classes. *CBE—Life Sciences Education*, 14(2), ar23. doi:10.1187/cbe.14-06-0095
- El Haddioui, I., & Khaldi, M. (2012). Learner behavior analysis on an online learning platform. *International Journal of Emerging Technologies in Learning (iJET)*, 7(2), 22-25.
- 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.
- Evertson, C. M. (1994). *Classroom management for elementary teachers*. Needham Heights, MA: Allyn & Bacon.
- Evertson, C. M., & Weinstein, C. S. (2006). Classroom management as a field of inquiry. *Handbook of classroom management: Research, Practice, and Contemporary Issues*, 3(1), 3-16.
- Flanders, N. A. (1961). Analyzing teacher behavior. *Educational leadership*, 19(3), 173-173.
- Froyen, L. A. (1988). *Classroom management: Empowering teacher-leaders*. Texas, TX: Merrill.
- Great Schools Partnership. (2014). *The Glossary of Education Reform*. Retrieved from <https://www.edglossary.org/classroom-management/>
- Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education*. Boston, MA: Center for Curriculum Redesign.
- Hsu, C. C., Chen, H. C., Su, Y. N., Huang, K. K., & Huang, Y. M. (2012). Developing a reading concentration monitoring system by applying an artificial bee colony algorithm to e-books in an intelligent classroom. *Sensors*, 12(10), 14158-14178.
- Huang, W., Li, N., Qiu, Z., Jiang, N., Wu, B., & Liu, B. (2020). An Automatic recognition method for students' classroom behaviors based on image processing. *Traitement du Signal*, 37(3), 503-509.
- Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Computers & Education: Artificial Intelligence*, 1, 100001. doi:10.1016/j.caeai.2020.100001
- Hwang, K. A., & Yang, C. H. (2009). Automated inattention and fatigue detection system in distance education for elementary school students. *Educational Technology & Society*, 12(2), 22-35.
- Jensen, F. V. (1996). *An Introduction to Bayesian networks* (Vol. 210, pp. 1-178). London, UK: UCL press.
- Kersting, N. (2008). Using video clips as item prompts to measure teachers' knowledge of teaching mathematics. *Educational and Psychological Measurement*, 68(5), 845-861.
- Kersting, N. B., Givvin, K. B., Thompson, B., Santagata, R., & Stigler, J. (2012). Developing measures of usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal*, 49(3), 568-590.
- Korpershoek, H., Harms, T., de Boer, H., van Kuijk, M., & Doolaard, S. (2016). A Meta-analysis of the effects of classroom management strategies and classroom management programs on students' academic, behavioral, emotional, and motivational outcomes. *Review of Educational Research*, 86(3), 643-680.
- Kounin, J. S. (1970). *Discipline and group management in classrooms*. New York, NY: Holt, Rinehart & Winston.
- Li, J., Tan, X., & Hu, Y. (2021). Research on the framework of intelligent classroom based on artificial intelligence. *The International Journal of Electrical Engineering & Education*, 0020720920984000. doi:10.1177/0020720920984000
- Li, S., Yan, M., Zhang, X., & Li, Z. (2020, November). Analysis on the application of AI technology in online education under the public epidemic crisis. In *International Conference on Innovative Technologies and Learning* (pp. 296-305). doi:10.1007/978-3-030-63885-6\_34
- Li, T. (2021). Research on intelligent classroom attendance management based on feature recognition. *Journal of Ambient Intelligence and Humanized Computing*, 1-8. doi:10.1007/s12652-021-03042-x

- O'Malley, K. J., Moran, B. J., Haidet, P., Seidel, C. L., Schneider, V., Morgan, R. O., Kelly, P. A., & Richards, B. (2003). Validation of an observation instrument for measuring student engagement in health professions settings. *Evaluation & the Health Professions*, 26(1), 86–103. doi:10.1177/0163278702250093
- Pourret, O., Naïm, P., & Marcot, B. (Eds.). (2008). *Bayesian networks: A Practical guide to applications*. New Jersey, NJ: John Wiley & Sons.
- Ramadan, R. A., Hagrass, H., Nawito, M., El Faham, A., & Eldesouky, B. (2010). The Intelligent classroom: Towards an educational ambient intelligence testbed. In *2010 Sixth International Conference on Intelligent Environments* (pp. 344-349). doi:10.1109/IE.2010.70
- Rich, P. J., & Hannafin, M. (2009). Video annotation tools: Technologies to scaffold, structure, and transform teacher reflection. *Journal of Teacher Education*, 60(1), 52-67.
- Schmidt, R. W. (1990). The Role of consciousness in second language learning. *Applied linguistics*, 11(2), 129-158.
- Tang, K. Y., Chang, C. Y., & Hwang, G. J. (2021). Trends in artificial intelligence-supported e-learning: A systematic review and co-citation network analysis (1998–2019). *Interactive Learning Environments*. doi:10.1080/10494820.2021.1875001
- Winer, L. R., & Cooperstock, J. (2002). The “Intelligent classroom”: Changing teaching and learning with an evolving technological environment. *Computers & Education*, 38(1-3), 253-266.
- Wu, J. Y., Liao, C. H., Cheng, T., & Nian, M. W. (2021). Using Data analytics to investigate attendees’ behaviors and psychological states in a virtual academic conference. *Educational Technology & Society*, 24(1), 75-91.
- Xenos, M. (2004). Prediction and assessment of student behaviour in open and distance education in computers using Bayesian networks. *Computers & Education*, 43(4), 345-359.
- Yang, B., Yao, Z., Lu, H., Zhou, Y., & Xu, J. (2020). In-classroom learning analytics based on student behavior, topic and teaching characteristic mining. *Pattern Recognition Letters*, 129, 224-231.
- Yang, M. T., Cheng, Y. J., & Shih, Y. C. (2011). Facial expression recognition for learning status analysis. In *International Conference on Human-Computer Interaction* (pp. 131-138). Berlin, Germany: Springer.
- Yang, S. J., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 2, 100008. doi:10.1016/j.caeai.2021.100008
- Zhu, Z. M., Xu, F. Q., & Gao, X. (2020). Research on school intelligent classroom management system based on internet of things. *Procedia Computer Science*, 166, 144-149.