The Journal of Educational Technology & Society has Impact Factor 2.133 and
5-Year impact factor 2.682 according to Thomson Scientific 2018 Journal Citations Report.

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

Jan. 2020 Journal of Educational Technology & Society

vol.23
no.1

Journal of
Educational Technology & Society

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

Volume 23 Issue 1
January 2020

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

Educational Technology & Society

An International Journal

Aims and Scope

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 inquires 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.

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

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

For Special Issue Proposal submission, please see:
https://www.j-ets.net/journal_info/special-issue-proposals

Submission procedure
All submissions must be uploaded through our online management system (https://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** (Subject: Submission for Educational Technology & Society journal).

Journal of Educational Technology & Society

Volume 23 Number 1 2020

Table of contents

Full Length Articles

The Effectiveness of the Flipped Classroom on Students' Learning Achievement and Learning Motivation: A Meta-Analysis

Lanqin Zheng[1], Kaushal Kumar Bhagat[2*], Yuanyi Zhen[1] and Xuan Zhang[1]
[1]School of Educational Technology, Faculty of Education, Beijing Normal University, Beijing, China // [2]Centre for Educational Technology, Indian Institute of Technology Kharagpur, India // bnuzhenglq@bnu.edu.cn // kkntnu@hotmail.com // zyyouc@126.com // zhxuan@mail.bnu.edu.cn
[*]Corresponding author

ABSTRACT: The purpose of this study was to examine the overall effectiveness of the flipped classroom on students' learning achievement and motivation. Data were collected from three databases, which include Web of Science, Scopus, and Eric. The present meta-analysis synthesized the findings of 95 studies with 15386 participants published from 2013 to 2019. The results revealed that the flipped classroom approach had a moderate effect size for learning achievement and learning motivation. The effect sizes of 12 moderators, including sample level, sample size, learning domain, the flip classroom model, research design, intervention duration, teaching method in the class, sample region, interactions in a pre-class and face-to-face class, tools in pre-class, and resources in pre-class were also analyzed. The results indicated that sample size, intervention durations, and sample regions significantly moderated the effect sizes. The findings of this study are discussed in-depth, together with the implications for practices on the use of the flipped classroom approach.

Keywords: Flipped classroom, Learning achievement, Learning motivation, Meta-analysis

1. Introduction

The flipped classroom has gained significant attention in recent years. It is also considered as an "inverted classroom" or "reversed instruction" (Bergmann & Sams, 2012). In the flipped classroom, learners watch the content videos at home and solve problems in the class (Tucker, 2012). The flipped classroom switches the in-class time and out-of-class time to enable more interactions between teachers and students in the class (Lai & Hwang, 2016). For example, Bergmann and Sams (2012) mentioned that in a traditional classroom, the main activities consisted of 5 minutes' warm-up activity, 20 minutes' review, 30 minutes' lecture, and 20 minutes' practice or lab activity. On the other hand, the activities of the flipped classroom include 5 minutes' warm-up activity, 10 minutes Q&A time on video, and 75 minutes of practice or lab activity. Class time is mainly used for collaboration among the students, discussion, and personalized learning (Francl, 2014).

Furthermore, several flipped models with different focuses have been proposed and implemented in practice. For example, the conventional flipped classroom emphasized content delivery (Bergmann & Sams, 2012). The FLIPPED model proposed by Chen, Wang, Kinshuk, and Chen (2014) advocated progressive activities, engaging experiences, and diversified platforms. These flipped classroom models are very promising and helpful for both research and practice. Previous studies found the positive effects of the flipped classroom and reported that the use of flipped classroom promoted students' learning performance (Lin, Hwang, Fu, & Chen, 2018) and learning satisfaction (Sergis, Sampson, & Pelliccione, 2018) compared to the traditional classroom. However, some studies indicated that there was no difference between the flipped classroom and the traditional classroom (Sparks, 2013; Strayer, 2012). So far, the effects of the flipped classroom are still debatable among the researchers. Therefore, it is vital to investigate the effects of the flipped classroom through systemic and comprehensive meta-analysis. This study aims to examine the effects of the flipped classroom and provide a clear picture about the mediating effects of moderator variables.

1.1. Previous reviews and meta-analysis of flipped classroom

The specifications for adopting the flipped classroom approach have been documented in previous literature reviews. For example, Seery (2015) analyzed the emerging trends on integrating the flipped learning model in chemistry in higher education. The findings indicated that the flipped learning approach developed an active learning environment

that resulted into a better conceptual understanding of learning engagement. Nederveld and Berge (2015) presented several tools for creating the flipped classroom in the workplace and discussed the benefits as well as challenges of the flipped classroom approach. O'Flaherty and Phillips (2015) conducted a systematic review of the flipped classroom in higher education. They found that the flipped classroom approach can improve academic performance and satisfaction. In another study, Kashada, Li, and Su (2017) analyzed ten studies related to the flipped classroom and examined the effects of the flipped classroom on students' performance in K-12 education. They found a positive impact of the flipped classroom on students' learning achievement. Njie-Carr, Ludeman, Lee, Dordunoo, Trocky, and Jenkins (2017) conducted a comprehensive review of relevant research concerning the flipped classroom model in nursing education. They provided the design and process information as well as the current status of the flipped classroom models through an analysis of 13 studies published in 2016. Lo and Hew (2017) conducted a literature review of the flipped classroom in K-12 education by analysis of 15 articles. They found that the flipped classroom model had a positive or neutral impact on learning achievement in K-12 education. However, some previous studies also reported the limitations of the flipped classroom. For example, Mellefont and Fei (2016) found that students' lack of preparation may hinder the effectiveness of the flipped classroom. Students were easily distracted when they watched the video (Toto & Nguyen, 2009). Besides, the effectiveness of the flipped classroom heavily relied on students' self-motivation (Wang, 2017). It is also difficult for teachers to monitor student comprehension and provide real-time feedback for each student (Milman, 2012).

Furthermore, some researchers conducted meta-analysis studies to examine the effectiveness of the flipped classroom. For example, Rahman et al. (2014) reviewed 15 studies on the flipped classroom. The results showed that the flipped classroom had a positive impact on students' achievement. The researchers conducted only qualitative analysis without calculating the effect size. Hew and Lo (2018) conducted a meta-analysis on 28 studies in the domain of health professionals and found a significant effect size in favor of the flipped classroom as compared to the traditional classroom. A meta-analysis study by Gillette et al. (2018) examined the effect of the flipped classroom in the pharmacy education and found a small positive effect for using the flipped model instead of the traditional lecture-based classroom. In the recent study, Cheng, Ritzhaupt, and Antonenko (2019) studied the overall effect of the flipped classroom approach. They also included different subject domains, student levels, and study durations as the moderator variables. The results indicated a moderate but significant positive effect of the flipped classroom on students' learning achievement.

1.2. The need for this study

Although previous reviews and meta-analysis analyzed the current status of the flipped classroom, there were three significant shortcomings of previous studies. First, very few meta-analysis studies have examined the effect of flipped classrooms compared to traditional classrooms on both learning achievement and learning motivation. Second, a systematic meta-analysis of the flipped classroom based on activity theory has not been published yet. Third, previous meta-analysis studies only analyzed the effects of three moderators or only focused on the specific subject domains, such as health professional or pharmacy education. There is a lack of a comprehensive meta-analysis to examine more moderators and cover all studies from 2013 to 2019. The present study is an attempt to fill the above research gaps.

1.3. Research questions

The purpose of the present study is twofold: the first aims to evaluate the effectiveness of the flipped classroom approach. Another is to examine whether moderator variables influence the effects of the flipped classroom on learning achievement. This study examined the effects of 12 moderator variables, including sample levels, sample size, learning domains, flip classroom models, research design, intervention durations, teaching methods in the class, sample regions, interactions in a pre-class and face-to-face class, tools in pre-class, and resources in pre-class. Therefore, the following research questions were proposed:

- What is the overall effectiveness of using the flipped classroom on students' learning achievement and learning motivation compared to the traditional classroom?
- How do various moderator variables influence on the effects of the flipped classroom?

2. Method

2.1. Data source

The data of this study were taken from three databases, including Web of Science, Scopus, and Eric. All of the studies relevant to the flipped classroom published from 2013 to 2019 were downloaded and further analyzed. Two sets of keywords were adopted to search research papers: (1) flipped classroom-related keywords, including flipped classroom, inverted classroom, flipped learning, flipped approach, and flipped-classroom; (2) learning achievement-related and learning motivation-related keywords, including learning outcome, learning achievement, achievement, outcome, learning performance, academic achievement, academic performance, learning motivation, motivation, and self-efficacy. The Boolean operator "AND" was adopted to integrate the two sets of keywords and the "OR" operator was used to connect within the set (Cooper, 2010).

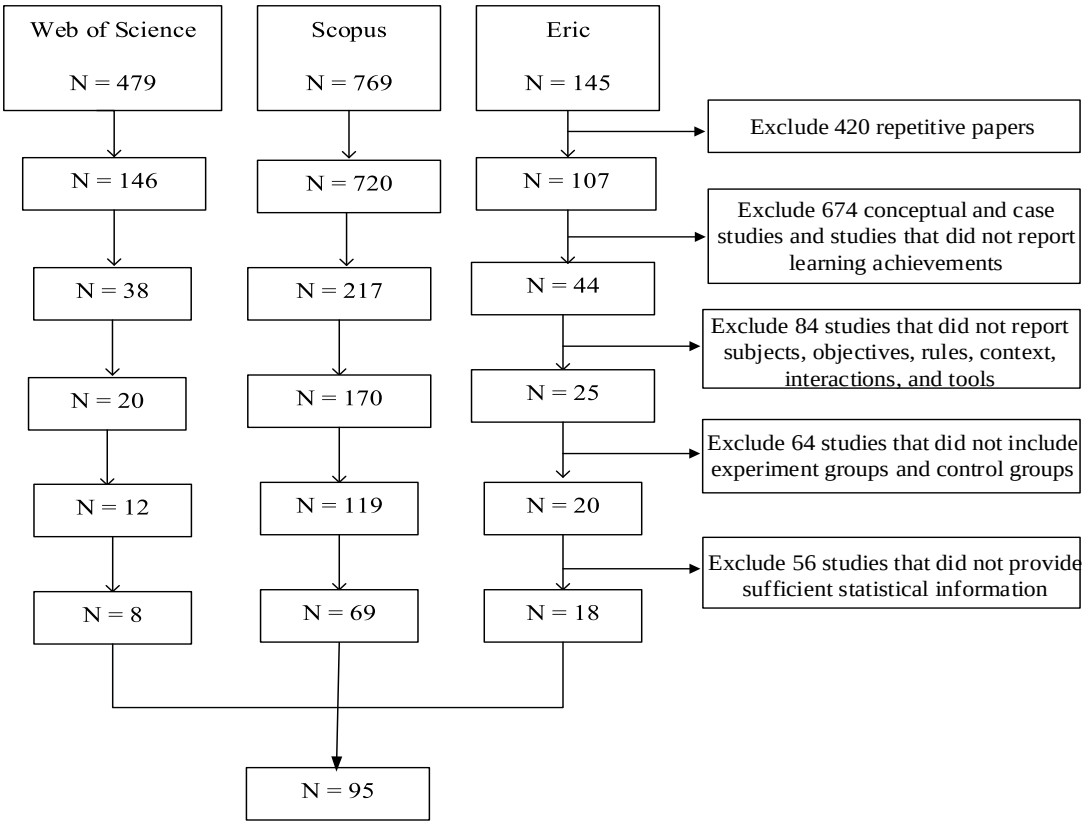


Figure 1. The search results

2.2. Search results

The research paper selection included two stages. The initial search yielded 1393 research papers, including 479 research papers from Web of Science, 769 research papers from Scopus, and 145 research papers from Eric. All of the research papers were examined, according to the following criteria:

- Studies published from 2013 to 2019.
- Research articles reported in English were only included in the present study. Studies not published in peer-reviewed journals (e.g., conference papers, book reviews, news, abstracts, and editorials) were excluded.
- The quasi-experimental or true-experimental studies were included. The conceptual studies were excluded. In addition, the selected studies should adopt the flipped classroom approach and report learning achievement and learning motivation.
- The selected studies should report how to implement the flipped classroom, including subjects, objectives, rules, context, interactions, and tools.

- The selected studies should include the experimental group and the control group. The studies should adopt the pretest to examine the equivalence of prior knowledge between the experimental and control groups. In addition, the instructors and learning content should be the same for the experimental and control groups.
- The selected studies should provide sufficient statistical information about learning achievement and motivation to calculate the effect size, such as means, standard deviations, *t* or *F* values, and the number of participants in each group.

Finally, 95 research papers were included in the present study for further analysis based on the above criteria. Figure 1 shows the search process and results.

2.3. Coding scheme

This study adopted activity theory as a model to analyze the features of the flipped classroom studies and the effects of moderator variables. Activity theory includes six components, including subject, object, mediating artifact, rules, community, and division of labor (Engeström, 1987). Engeström (2001) believed that activity theory represented the elements of learning activities and how learning activities occur. Moreover, previous studies also adopted activity theory to analyze different learning activities (Chung, Hwang, & Lai, 2019; Zheng et al., 2019). Figure 2 shows the adapted framework based on previous studies (Engeström, 1987; Sung, Yang, & Lee, 2017), and it includes six elements: subjects, objectives, rules, context, communication, and tools.

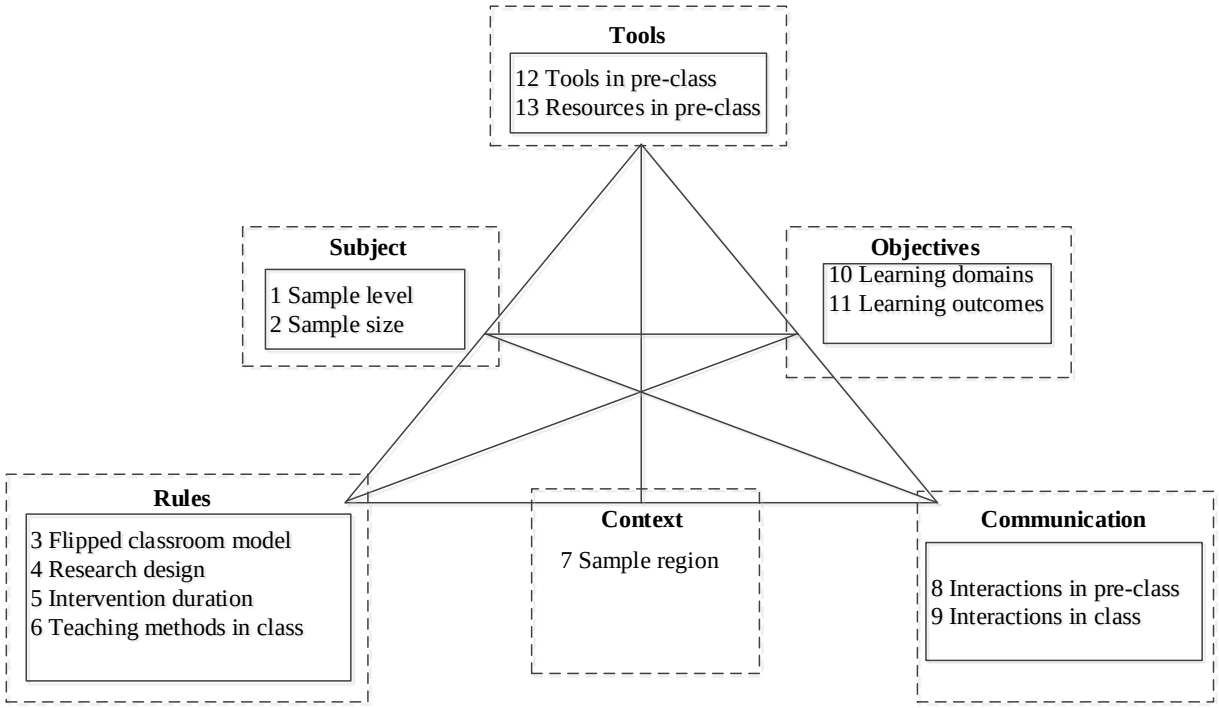


Figure 2. The analysis framework for flipped classroom based on activity theory

Table 1 shows the coding scheme in detail. Regarding learning outcomes, it includes learning achievement and learning motivation. Learning achievement is usually measured by standardized, teacher-made, or research-made tests to evaluate learners' knowledge acquisition or utilization (Sung, Yang, & Lee, 2017). Learning motivation was conceptualized as an established pattern of pursuing goals, beliefs, and emotions (Ford, 1992). In addition, the flipped classroom model included the traditional flipped classroom model and the innovative flipped classroom model. The traditional flipped classroom model refers to a teaching strategy that reverses what is done inside the classroom and outside the classroom (Abeysekera & Dawson, 2015). Innovative flipped classroom model refers to a new teaching strategy that integrating traditional flipped classroom model into other learning approaches such as social inquiry learning approach, problem-based learning, and so on. The coding scheme was developed based on the studies conducted by Zheng (2016), Zheng et al. (2019), and Sung, Chang, & Liu (2016). The coding process

included three steps proposed by Cooper (2010). First, three coders achieved a consensus about the definition of all entries by analyzing two papers. Second, three coders selected ten papers, independently coded, and negotiated until they achieve the consensus. Third, all of the rest papers were analyzed by three coders. The inter-coder Kappa reliability was 0.91.

Table 1. The coding scheme

Super-dimensions	Sub-dimensions	Coding scheme
Subjects	Sample level	(1) Primary school; (2) Junior and Senior High School; (3) Higher education.
	Sample size	(1) 1–50; (2) 51–100; (3) 101–300; (4) More than 300.
Objectives	Learning domain	(1) Natural Science (including science, mathematics, physics, biology, geography); (2) Social Science (including politics, education, psychology, linguistics); (3) Engineering & Technological Science (including engineering, computer science, educational technology); (4) Medical Science (including health and medicine).
	Learning outcome	Learning achievement; Learning motivation.
Rules	Flipped classroom model	(1) Traditional flipped classroom model; (2) Innovative flipped classroom model (e.g., technology enhanced flipped classroom, "Flipped" social inquiry learning model, clicker-aided flipped classroom, modern flipped classroom model, partial flipped classroom, flipped-blended classroom, in-flipped classroom, problem-based learning with flipped classroom).
	Research design	(1) True experimental design; (2) Quasi-experimental design.
	Intervention duration	(1) 2-4 weeks; (2) 5-8 weeks; (3) 9-24 weeks; (4) More than 24 weeks.
	Teaching method in F2F class	(1) One teaching method (e.g., problem-based learning or collaborative learning or self-directed study); (2) Two teaching methods (e.g., project-based learning and collaborative learning); (3) Three or more than three types of teaching methods (e.g., problem-based learning, collaborative learning, and inquiry-based learning).
Context	Sample region	(1) Africa; (2) Asia; (3) Europe; (4) North America; (5) Mixed region (e.g., China and US)
Communication	Interaction in pre-class	(1) Reading learning materials (one kind of interaction); (2) Watching the teaching videos (one kind of interaction); (3) Two types of interactions (e.g., watching the teaching videos, reading materials); (4) Three or more than three types of interactions (e.g., watch videos, reading learning materials, self-test)
	Interaction in F2F class	(1) Two types of interactions (e.g., group discussion and problem solving); (2) Three or more than three types of interactions (e.g., group discussion, presentation, and quiz).
Tools	Tools after class	(1) Online learning platform; (2) Others (Online discussion forum or game).
	Resources after class	(1) Video recordings; (2) Two types of resources (e.g., video recordings, readings); (3) Three or more than three types of resources (e.g., lectures, readings, video recordings).

2.4. Effect size calculation

The effect size calculation included four steps proposed by Borenstein, Hedges, Higgins, and Rothstein (2009). First, calculate the effect size of each study. Second, integrate the effect sizes of all studies to compute the overall weighted mean effect size by Hedges's *g*. Third, calculate the confidence interval for the overall mean effect size by the random effect model. Fourth, examine whether the moderator variables influenced the effect size through the Q_B value. A random-effect model was adopted to examine the impacts of moderator variables. The effect size was calculated using the Comprehensive Meta-analysis software. The publication bias was examined by the classic fail-safe N and Orwin's fail-safe N (Rosenthal, 1979). If the fail-safe N is above 5n+10 (n represents the number of studies), then it is unlikely to influence the effect size by the unpublished studies.

3. Results

3.1. Descriptive information

The present study analyzed the demographics of 95 studies and the features of the flipped classroom. The following sections will describe the results in detail. Table 2 presents the descriptive information of moderator variables and their percentages. There were 95 articles with 15,386 participants. With respect to subjects, the largest proportion of studies selected higher education and 51-300 participants. With regard to objectives, the most frequently selected learning domains were social science, followed by natural science and engineering as well as technological science. In terms of rules, most of the studies adopted quasi-experimental design to conduct studies for 9-24 weeks using the traditional flipped classroom model and two types of teaching methods. As for context, most studies implemented the flipped classroom in North America, followed by Asia. Concerning communication, most studies engaged participants in two types of interactions in pre-class and three or more than three types of interactions in class. Regarding tools, most studies adopted the online learning platform and two types of resources in the pre-class.

Table 2. The moderator variables categories and proportion of 95 studies

Variable	Category	No. of studies (k)	Proportion of studies
Sample levels	(1) Primary school	3	3.16%
	(2) Junior and Senior High School	14	14.74%
	(3) Higher education	78	82.1%
Sample size	(1) 1-50	13	13.69%
	(2) 51-100	36	37.89%
	(3) 101-300	36	37.89%
	(4) More than 300	10	10.53%
Subject domains	(1) Natural Science	30	31.58%
	(2) Social Science	34	35.79%
	(3) Engineering and Technological Science	16	16.84%
	(4) Medical Science	15	15.79%
Learning outcomes	(1) Learning achievements	95	100%
	(2) Learning motivation	9	9.47%
Research design	(1) Quasi-experimental design	90	94.74%
	(2) True experimental design	5	5.26%
Intervention durations	(1) 2-4 weeks	7	7.37%
	(2) 5-8 weeks	13	13.68%
	(3) 9-24 weeks	57	60.00%
	(4) More than 24 weeks	18	18.95%
Flipped classroom models	(1) Traditional flipped classroom model	81	85.26%
	(2) Innovative flipped classroom model	14	14.74%
Teaching methods in F2F class	(1) One teaching method	20	21.05%
	(2) Two teaching methods	60	63.16%
	(3) Three or more than three types of teaching methods	15	15.79%
Sample regions	(1) Africa	3	3.16%
	(2) Asia	42	44.21%

	(3) North America	43	45.26%
	(4) Europe	6	6.32%
	(5) Mixed region	1	1.05%
Interactions in pre-class	(1) Reading learning materials	5	5.26%
	(2) Watching the teaching videos	23	24.21%
	(3) Two types of interactions	43	45.27%
	(4) Three or more than three types of interactions	24	25.26%
Interactions in F2F class	(1) Two types of interactions	29	30.53%
	(2) Three or more than three types of interactions	66	69.47%
Tools in pre-class	(1) Online learning platform	89	93.68%
	(2) Others (Online discussion forum or game).	6	6.32%
Resources in pre-class	(1) Videos	25	26.32%
	(2) Two types of resources	49	51.57%
	(3) Three or more than three types of resources	21	22.11%

3.2. Overall effect size

Table 3 and Table 4 shows the overall effect sizes for learning achievement and learning motivation respectively. Based on the procedure of Borenstein et al. (2009), a random effect model was adopted to calculate the effect sizes of 95 selected studies. The results indicated that the overall effect size for learning achievement was 0.663, with a 95% confidence interval of 0.544-0.783. The effect sizes of 0.80, 0.50, and 0.20 were regarded as a larger, medium, and small effect size respectively based on Cohen's (1992) finding. Therefore, the flipped classroom approach had a medium effect size on students' learning achievement. The test of heterogeneity revealed that the effect sizes were heterogeneous in the present study (Q_{total} = 1192.145, z = 10.877, p < 0.001). In addition, the flipped classroom approach had a medium effect size on students' learning motivation (ES = 0.661). The results of heterogeneity analysis indicated the effect sizes were heterogeneous in this study (Q_{total} = 70.95, z = 2.999, p < 0.005). These findings also revealed that the significant differences among the effect sizes were due to sources other than subject-level sample error (Sung, Yang, & Lee, 2017).

Table 3. Overall effect sizes of learning achievement

	k	ES	*SE*	σ^2	95% CI		Test of mean		Test of heterogeneity		
					Lower	Upper	*Z*	*p*	*Q*	*df(Q)*	*p*
Fixed	95	0.501	0.016	0.000	0.468	0.533	30.473	.000	1192.145	94	.000
Random	95	0.663	0.061	0.004	0.544	0.783	10.877	.000			

Table 4. Overall effect sizes of learning motivation

	k	ES	*SE*	σ^2	95% CI		Test of mean		Test of heterogeneity		
					Lower	Upper	*Z*	*p*	*Q*	*df(Q)*	*p*
Fixed	9	0.437	0.071	0.005	0.299	0.576	6.196	.000	70.950	8	.000
Random	9	0.661	0.220	0.049	0.229	1.093	2.999	.003			

3.3. Effect sizes of learning achievements for moderator variables

The random-effect model was adopted to analyze the effect size of each moderator variable. Table 5 shows the results of twelve moderator variables.

3.3.1. Subjects

It was found that the flipped classroom studies implemented in junior and senior high school produced the largest effect size, followed by higher education and primary school. However, Q_B did not achieve statistical significance. Regarding the sample size, it was found that the sample size of 1-50 produced the largest effect size, followed by 51-100, 101-300, and more than 300. In addition, Q_B reached statistical significance (Q_B = 11.290, df = 3, p = .010), showing that the effect sizes of different sample sizes differed significantly.

3.3.2. Objectives

Table 5 demonstrated that the effect size for natural science domain achieved the highest effect size, followed by engineering & technological science, medical science, and social science. However, Q_B did not achieve statistical significance, which means that there was no significant difference among different subject domains.

3.3.3. Rules

Table 5 also indicated that the traditional flipped classroom model produced a larger effect size than the innovative flipped classroom model. However, the test of heterogeneity indicated that there was no significant difference between the two types of flipped classroom models. With respect to research design, the findings revealed that the true experimental design had a higher effect size and the quasi-experimental design had the lowest effect size. Both the two types of research design showed significant effect sizes. However, the Q_B did not achieve the significance, showing that the average effect sizes did not significantly differ between the true-experimental and quasi-experimental design.

With regard to the intervention duration, the findings indicated that interventions of 5-8 weeks had the largest effect size, followed by interventions of 2-4 weeks, interventions of 9-24 weeks, interventions of more than 24 weeks. Additionally, the Q_B was significant (Q_B = 9.458, df = 3, p = .024), which suggested that the average effect size differed significantly within the four types of intervention durations.

In terms of teaching methods in a face-to-face classroom, the results indicated that one teaching method had the largest effect size, followed by three or more than three types of teaching methods, and two teaching methods. However, the Q_B did not achieve the significance, showing that the average effect sizes did not significantly differ among different types of teaching methods.

3.3.4. Context

Table 5 indicated that the flipped classroom approach produced the largest effect size in Africa, followed by mixed region, Asia, Europe, and North America. The test of heterogeneity indicated that there was a significant difference among five types of sample regions (Q_B = 21.066, df = 4, p = .000).

3.3.5. Communications

This study analyzed two types of communications for flipped classroom studies. One was interaction in pre-class and another was interaction within class. It was found that watching the teaching videos yielded the largest effect size, followed by reading learning materials, two types of interactions, and three or more than three types of activities. However, the Q_B did not achieve the significance, showing that the average effect sizes did not significantly differ among different types of interactions. Concerning interactions in a face-to-face class, the results indicated that two types of interactions had the highest effect size and three or more than three types of interactions had the lowest effect size. However, the Q_B did not achieve the significance, showing that the average effect sizes did not significantly differ.

3.3.6. Tools

It was found that online discussion forum or game produced a larger effect size than the online learning platform. The test of heterogeneity indicated that there was no significant difference between the two types of tools in pre-class. In addition, the data given in Table 5 demonstrated that the effect size for video recordings achieved the highest effect size, followed by three or more than three types of resources. However, the Q_B did not achieve the significance, showing that the average effect sizes did not differ significantly.

Table 5. The analysis results for moderator variables

Category	*k*	*g*	*z*	95% CI	Q_B	*df*
Sample levels					0.828	2
1. Primary school	3	0.541	1.515	[-0.159,1.241]		
2. Junior and Senior High School	14	0.793	4.905***	[0.476,1.110]		
3. Higher education	78	0.646	9.576***	[0.513,0.778]		
Sample size					11.290**	3
1. 1-50	13	0.953	5.340***	[0.603,1.303]		
2. 51-100	36	0.830	8.250***	[0.633,1.028]		
3. 101-300	36	0.534	5.623***	[0.348,0.720]		
4. More than 300	10	0.312	1.781	[-0.031,0.655]		
Learning domains					1.266	3
1. Engineering& Technological Science	16	0.693	4.635***	[0.400,0.985]		
2. Medical Science	15	0.662	4.525***	[0.375,0.948]		
3. Natural Science	30	0.740	6.948***	[0.531,0.948]		
4. Social Science	34	0.576	5.602***	[0.375,0.778]		
Flipped classroom models					0.405	1
1. Traditional flipped classroom model	81	0.680	10.230***	[0.550,0.810]		
2. Innovative flipped classroom model	14	0.570	3.567***	[0.257,0.883]		
Research design					0.240	1
1. True experimental design	5	0.793	2.918**	[0.260,1.326]		
2. Quasi-experimental design	90	0.657	10.459***	[0.534,0.780]		
Intervention Durations					9.458*	3
1. 2-4 weeks	7	0.774	3.322***	[0.317,1.230]		
2. 5-8 weeks	13	1.112	6.439***	[0.774,1.451]		
3. 9-24 weeks	57	0.626	7.893***	[0.471,0.781]		
4. More than 24 weeks	18	0.458	3.346***	[0.190,0.726]		
Teaching methods in F2F class					4.753	2
1. One teaching method	20	0.891	6.721***	[0.631,1.150]		
2. Two teaching methods	60	0.570	7.561***	[0.422,0.717]		
3. Three or more than three types of teaching methods	15	0.743	4.785***	[0.439,1.047]		
Sample regions					21.066***	4
1. Africa	3	1.352	3.725***	[0.641,2.063]		
2. Asia	42	0.913	10.013	[0.734,1.091]		
3. Europe	6	0.627	2.668**	[0.166,1.088]		
4. North America	43	0.397	4.638***	[0.229,0.565]		
5. Mixed region	1	0.993	1.667	[0.734,1.091]		
Interactions in pre-class					2.712	3
1. Reading learning materials	5	0.698	2.707**	[0.193,1.202]		
2. Watching the teaching videos	23	0.774	6.349***	[0.535,1.014]		
3. Two types of interactions	43	0.685	7.733***	[0.512,0.859]		
4. Three or more than three types of interactions	24	0.504	4.238***	[0.271,0.737]		
Interactions in F2F class					0.109	1

1. Two types of interactions	29	0.695	6.190***	[0.475,0.915]		
2. Three or more than three types of interactions	66	0.651	8.839***	[0.506,0.795]		
Tools in pre-class					0.123	1
1. Online learning platform	89	0.658	10.448***	[0.534,0.781]		
2. Others	6	0.746	3.058**	[0.268,1.224]		
Resources in pre-class					3.521	2
1. Video recordings	25	0.838	7.308***	[0.614,1.063]		
2. Two types of resources	49	0.576	7.150***	[0.418,0.734]		
3. Three or more than three types of resources	21	0.647	5.201***	[0.403,0.891]		

Note. $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

3.4. Publication bias

The publication bias was evaluated by the funnel plot, the classic fail-safe N, and Orwin's fail-safe N. As shown in Figure 3, it was found that the funnel plot had symmetrical distribution. Therefore, there was no publication bias in the present meta-analysis. As shown in Table 6, the results of the classic fail-safe N indicated that 4885 missing studies would be needed to nullify the effect size, which was far larger than 485 (5n+10). Furthermore, the result of Orwin's fail-safe N revealed that 4662 missing studies would be needed to reduce Hedges's g to a trivial level (see Table 7). Therefore, the findings indicated that this meta-analysis was not affected by publication bias.

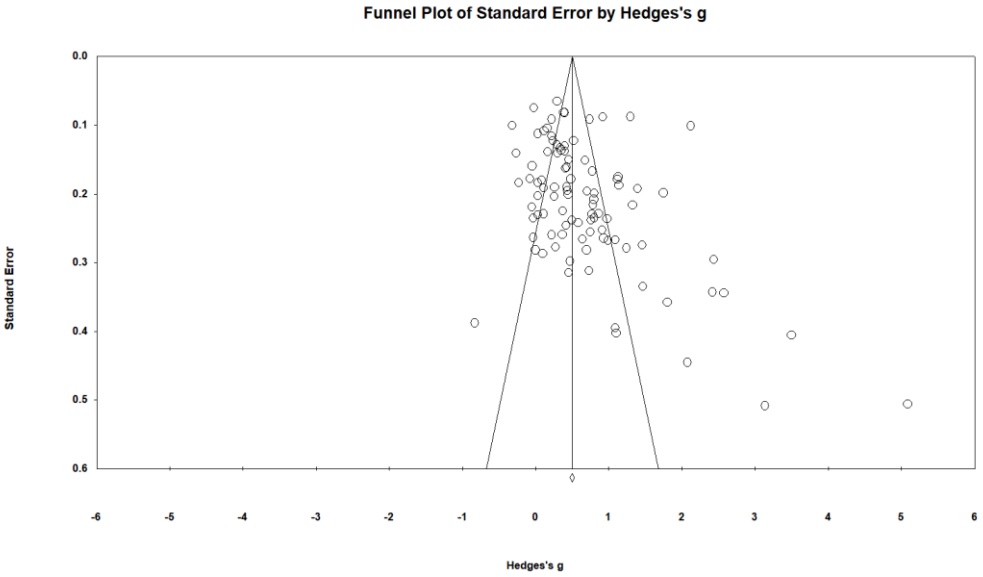


Figure 3. Funnel plot of standard error by effect size

Table 6. Classic fail-safe N

Items	Value
Z value for observed studies	31.845
p value for observed studies	0.000
Alpha	0.050
Tails	2.000
Z for alpha	1.960
Number of observed studies	95
Number of missing studies that would bring p value to > alpha	4885

Table 7. Orwin's fail-safe N

Items	Value
Hedges's g in observed studies	0.501
Criterion for a 'trivial' Hedges's g	0.010
Mean Hedges's g in missing studies	0.000
No. of missing studies needed to reduce Hedges's g to <0.01	4662

4. Discussion

This study examined the effects of the flipped classroom approach on students' learning achievement and learning motivation compared to traditional lecture-based instruction. Based on a total of 95 eligible studies with a total of 15386 students, it was found that the flipped classroom approach had an overall positive effect on students' learning achievement and learning motivation. The finding expanded the previous studies and revealed that the use of the flipped classroom had a significant impact on learning motivation through a comprehensive meta-analysis. The present study also provided substantial evidence on how the use of the flipped classroom was moderated by 12 variables, including sample levels, sample size, learning domains, flip classroom models, research design, intervention durations, teaching methods in the class, sample regions, interactions in a pre-class and face-to-face class, tools in pre-class, and resources in pre-class.

4.1. Sample level and sample size

For the sample level, it was found that there was no significant difference among the three sample levels. This result was similar to the findings of Cheng et al. (2019) in which they did not find any significant effect of sample level in the flipped classroom. The studies conducted at the junior and senior high school showed larger effects as compared to higher education. There was no significant effect size for the primary level. This may be because very few studies used the primary level as the sample for their study. Furthermore, this study also found that the small sample size had the largest effect size. The main reason was that the small sample size produced the less variation source, which led to the larger effect size (Slavin & Smith, 2009).

4.2. Learning domains

For learning domains, there was no significant difference among different subject domains. This finding indicated that learning domains did not have a significant impact on the effectiveness of the flipped classroom. This result might be explained by the fact that the appropriate use of flipped classroom would be effective in any learning domains that include real-world problems, design effective in-class learning activities, facilitate efficient interactions through information technologies, and integrate other pedagogical models according to the characteristics of different learning domains. Furthermore, natural science, engineering and technological science, medical science, and social science showed positive and medium effects size. However, Cheng et al. (2019) found that there was a significant difference in learning domains. The possible reason could be that the data sources and statistical information were different between the two studies.

4.3. Interventions

The findings revealed that there was no significant difference in the flipped classroom models. Therefore, the practitioners can select either the traditional flipped classroom or innovative classroom model. In addition, it was found that 94.7% of the studies in the present meta-analysis employed quasi-experimental design, and only 5.3% selected true experiments. The effect size of the true experimental design was larger than the quasi-experimental design. Therefore, more true-experimental studies need to be conducted in the flipped classroom research. Moreover, the present study revealed that the medium intervention duration produced the largest effect size. The main reason might be that too long durations will produce potential variation, and too short durations cannot validate the effectiveness of the flipped classroom. In terms of teaching methods in a face-to-face classroom, it was found that

there was no significant difference among different types of teaching methods. Thus, teachers and practitioners can select appropriate teaching methods based on instructional objectives and content.

4.4. Sample regions

The results indicated that there was a significant difference among the five types of sample regions. The studies conducted in the Africa region showed the most significant effects of the flipped classroom. The reason may be that the flipped classroom model is helping the developing countries to enhance the learning achievements and motivation of the students, which is a very significant output.

4.5. Interaction types

The findings revealed that different types of interactions in pre-class and the face-to-face class did not differ significantly. However, it was found that watching the teaching videos yielded the largest effect size. This could be explained that watching teaching videos is very important for a better understanding of learning content in a flipped classroom. Therefore, it is strongly recommended to develop high-quality videos that include recordings with elaborated instructional design, clear pictures, content-rich learning materials, and high-degree interactions, which can engage the learners prior to class.

4.6. Tools

The results indicated that the online discussion forum or game produced a larger effect size than the traditional online learning platform. The reason may be that the effective application of advanced technologies in the flipped classroom can promote learning achievement and learning motivation (Lin, 2019). In terms of resources in pre-class, it was found that video recordings achieved the highest effect size. Therefore, it is recommended to develop high-quality video recordings to facilitate the flipped classroom.

4.7. Implications

The present study has several implications for implementation of the flipped classroom, which are described and analyzed below.

4.7.1. Enhancing the research design quality for the flipped classroom interventions

The present meta-analysis found that different sample regions, sample sizes, and intervention durations had significant impacts on effect size. In order to enhance the research design quality, the following aspects may be considered by researchers and practitioners before the flipped classroom implementation. First, the characteristics of participants should be taken into account before the implementation of the flipped classroom approach. Students' experiences, prior knowledge, information and communication technology skills, and attitude toward the flipped classroom had great impacts on the effectiveness of the flipped classroom. Furthermore, if the participants come from mixed regions, their cultural background is another important factor for the flipped classroom interventions.

Second, the present meta-analysis indicated that less than 300 participants could produce a large effect size. The largest effect size was produced by less than 50 participants in this study. Previous studies reported that the appropriate sample size could ensure unbiased findings and estimates (McNeish & Stapleton, 2016). Therefore, it is suggested that the sample size should be less than 300 participants to decrease the potential variation source.

Third, the midterm intervention duration is more appropriate than shorter or longer intervention duration. Previous studies revealed that the intervention duration affected the reliability and validity of the research (Sung, Chang, & Liu, 2016). The present meta-analysis found that the midterm intervention duration (5-8 weeks) produced the largest effect size. It is very difficult to yield any effects for a too short duration. In addition, it will take lots of time to

introduce long-term flipped classroom implementation. Therefore, the teachers and practitioners may adopt the midterm intervention duration to implement the flipped classroom.

4.7.2. Integrating other pedagogical models with the flipped classroom approach

The appropriate pedagogy can improve the effectiveness of the flipped classroom. The traditional flipped classroom ignored the activity delivery and students' experiences (Chen, Wang, Kinshuk, & Chen, 2014). By integrating other pedagogical models such as collaborative learning, inquiry-based learning, and problem-based learning into the flipped classroom, the effectiveness of the flipped classroom can be maximized. These pedagogical models included modified flipped classroom (Scott, Green, & Etheridge, 2016), flipped social inquiry learning approach (Jong, 2017), clicker-aided flipped classroom (Yu & Wang, 2016), in-flipped classroom (Chiang, & Wang, 2015), problem-based learning with the flipped classroom (Tsai, Shen, & Lu, 2015), and so on. Therefore, it is suggested that educators and practitioners can harness an innovative pedagogy to implement the flipped classroom.

5. Conclusions

This meta-analysis provided substantial evidence for the positive effect of adopting the flipped classroom and how those effects were influenced by different moderator variables. The main findings were summarized as below:

- The flipped classroom approach had a moderate effect size of 0.663 for learning achievement and a moderate effect size of 0.661 for learning motivation.
- The results indicated that sample size, intervention durations, and sample regions significantly moderated the effect sizes.
- The small sample size (1-50) had a larger effect size than the large sample size (more than 50).
- The true experimental design had better effects than the quasi-experimental design.
- The midterm interventions (5-8 weeks) produced better effects than short duration (shorter than five weeks) and long-term intervention duration (longer than eight weeks).
- Watching the teaching videos yielded the largest effect size in pre-class, and videos produced a better effect size than other resources in pre-class.

These findings are very promising and provide insight into the implementation of the flipped classroom in the future. However, this study has several limitations. First, due to the limited empirical studies on the flipped classroom approach, only 95 empirical studies reported sufficient statistical information and descriptive information about the flipped classroom. In the future study, the data source, including grey literature and unpublished studies, needs to be expanded further to get a more comprehensive understanding of the flipped classroom. Second, the present study analyzed the effects of 12 moderators. Further studies are needed to explore the effect sizes of other moderators. For example, achievement indicators (standardized achievement test or self-reported grades) are also a source of variance that can be analyzed as a moderator. Finally, this study only analyzed the effects of the flipped classroom approach on learning achievement and learning motivation. Future studies may examine the effects of the flipped classroom approach on other dependent variables, such as learning behavior or learning attitude.

Acknowledgement

This study is funded by the youth project of Humanities and Social Science Research in the Ministry Education (19YJC880141) and National Natural Science Foundation of China (61907003).

References

Abeysekera, L., & Dawson, P. (2015). Motivation and cognitive load in the flipped classroom: Definition, rationale and a call for research. *Higher Education Research and Development*, *34*, 1–14. doi:10.1080/07294360.2014.934336

Bergmann, J., & Sams, A. (2012). *Flip your classroom: Reach every student in every class every day*. Washington, DC: Internal Society for Technology in Education.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Chen, Y., Wang, Y., Kinshuk, & Chen, N. S. (2014). Is FLIP enough? Or should we use the FLIPPED model instead? *Computers & Education*, *79*, 16–27. doi:10.1016/j.compedu.2014.07.004

Cheng, 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*(4), 793–824. doi:10.1007/s11423-018-9633-7

Chiang, Y. H., & Wang, H. C. (2015). Effects of the in-flipped classroom on the learning environment of database engineering. *International Journal of Engineering Education*, *31*(2), 454–460.

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. doi:10.1016/j.compedu.2018.10.010

Cohen, J. (1992). A Power primer. *Psychological Bulletin*, *112*, 155–159. doi:10.1037/0033-2909.112.1.155

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). London, UK: Sage Publications.

Engeström, Y. (1987). *Learning by expanding: An activity-theoretical approach to developmental research*. Helsinki, Finland: Prienta-Konsultit Oy.

Engeström, Y. (2001). Expansive learning at work: Toward an activity-theoretical reconceptualization. *Journal of Education and Work*, *14*(1), 133–156. doi:10.1080/13639080020028747

Ford, M. E. (1992). *Human motivation: Goals, emotions, and personal agency beliefs*. Newbury Park, CA: Sage.

Francl, T. J. (2014). Is flipped learning appropriate. *Journal of Research in Innovative Teaching, 71*, 119–128.

Gillette, 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), 6898–6898. doi:10.5688/ajpe6898

Hew, K. F., & Lo, C. K. (2018). Flipped classroom improves student learning in health professions education: A Meta-analysis. *BMC Medical Education, 18*: 38. doi:10.1186/s12909-018-1144-z

Jong, M. S. Y. (2017). Empowering students in the process of social inquiry learning through flipping the classroom. *Educational Technology & Society, 20*(1), 306–322.

Kashada, A., Li, H., & Su, C. (2017). Adoption of flipped classrooms in K-12 education in developing countries: Challenges and obstacles. *International Journal of Emerging Technologies in Learning*, *12*(10), 147–157. doi:10.3991/ijet.v12i10.7308

Lai, C.-L., & Hwang, G.-J. (2016). A Self-regulated flipped classroom approach to improving students' learning performance in a mathematics course. *Computers & Education, 100*, 126–140. doi:10.1016/j.compedu.2016.05.006

Lin, C. J., Hwang, G. J., Fu, Q. K., & Chen, J. F. (2018). A Flipped contextual game-based learning approach to enhancing EFL students' English business writing performance and reflective behaviors. *Journal of Educational Technology & Society*, *21*(3), 117–131.

Lin, Y. T. (2019). Impacts of a flipped classroom with a smart learning diagnosis system on students' learning performance, perception, and problem solving ability in a software engineering course. *Computers in Human Behavior*, *95*, 187–196. doi:10.1016/j.chb.2018.11.036

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

McNeish, D. M., & Stapleton, L. M. (2016). The Effect of small sample size on two-level model estimates: A Review and illustration. *Educational Psychology Review, 28*(2), 295–314. doi:10.1007/s10648-014-9287-x

Mellefont, L. A., & Fei, J. G. (2016). Student perceptions of 'flipped' microbiology laboratory classes. *International Journal of Innovation in Science and Mathematics Education*, *24*(1), 24–35.

Milman, N. (2012). The Flipped classroom strategy: What is it and how can it be used? *Distance Learning, 9*(3), 85–87.

Nederveld, A., & Berge, Z. L. (2015). Flipped learning in the workplace. *Journal of Workplace Learning*, *27*(2), 162–172. doi:10.1108/JWL-06-2014-0044

Njie-Carr, V. P., Ludeman, E., Lee, M. C., Dordunoo, D., Trocky, N. M., & Jenkins, L. S. (2017). An Integrative review of flipped classroom teaching models in nursing education. *Journal of Professional Nursing*, *33*(2), 133–144. doi:10.1016/j.profnurs.2016.07.001

O'Flaherty, J., & Phillips, C. (2015). The Use of flipped classrooms in higher education: A Scoping review. *The Internet and Higher Education*, *25*(4), 85–95. doi:10.1016/j.iheduc.2015.02.002

Rahman, A. A., Aris, B., Mohamed, H., & Zaid, N. M. (2014). The Influences of flipped classroom: A Meta-analysis. In *Proceedings of IEEE 6th Conference on Engineering Education (ICEED)* (pp. 24–28). doi:10.1109/ICEED.2014.7194682

Rosenthal, R. (1979). The File drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. doi:10.1037/0033-2909.86.3.638

Scott, C. E., Green, L. E., & Etheridge, D. L. (2016). A Comparison between flipped and lecture-based instruction in the calculus classroom. *Journal of Applied Research in Higher Education*, *8*(2), 252–264. doi:10.1108/JARHE-04-2015-0024

Seery, M. K. (2015). Flipped learning in higher education chemistry: emerging trends and potential directions. *Chemistry Education Research and Practice*, *16*(4), 758–768. doi:10.1039/C5RP00136F

Sergis, S., Sampson, D. G., & Pelliccione, L. (2018). Investigating the impact of Flipped Classroom on students' learning experiences: A Self-Determination Theory approach. *Computers in Human Behavior*, *78*, 368–378. doi:10.1016/j.chb.2017.08.01

Slavin, R. E., & Smith, D. (2009). Effects of sample size on effect size in systematic reviews in education. *Educational Evaluation and Policy Analysis*, *31*(4), 500–506. doi:10.3102/0162373709352369

Sparks, R. J. (2013). Flipping the classroom: An Empirical study examining student learning. *Journal of Learning in Higher Education*, *9*(2), 65–70.

Strayer, J. F. (2012). How learning in an inverted classroom influences cooperation, innovation and task orientation. *Learning environments research*, *15*(2), 171–193. doi:10.1007/s10984-012-9108-4.

Sung, Y. T., Chang, K. E., & Liu, T. C. (2016). The Effects of integrating mobile devices with teaching and learning on students' learning performance: A Meta-analysis and research synthesis. *Computers & Education*, *94*, 252–275. doi:10.1016/j.compedu.2015.11.008

Sung, Y. T., Yang, J. M., & Lee, H. Y. (2017). The Effects of mobile-computer-supported collaborative learning: Meta-analysis and critical synthesis. *Review of Educational Research*, *87*(4), 768–805. doi:10.3102/0034654317704307.

Toto, R., & Nguyen, H. (2009). Flipping the work design in an industrial engineering course. In *Proceedings of 39th IEEE Frontiers in Education Conference* (pp. 1–4). San Antonio, TX: IEEE. doi:10.1109/FIE.2009.5350529.

Tsai, C. W., Shen, P. D., & Lu, Y. J. (2015). The Effects of problem-based learning with flipped classroom on elementary students' computing skills: A case study of the production of ebooks. *International Journal of Information and Communication Technology Education*, *11*(2), 32–40. doi:10.4018/ijicte.2015040103

Tucker, B. (2012). The flipped classroom. *Education Next*, *12*(1), 82–83.

Wang, T. (2017). Overcoming barriers to 'flip': Building teacher's capacity for the adoption of flipped classroom in Hong Kong secondary schools. *Research and practice in technology enhanced learning*, *12*(1), 6. doi:10.1186/s41039-017-0047-7

Yu, Z., & Wang, G. (2016). Academic achievements and satisfaction of the clicker-aided flipped business English writing class. *Journal of Educational Technology & Society*, *19*(2), 298–312.

Zheng, L., Chen, N. S., Cui, P., & Zhang, X. (2019). A Systematic review of technology-supported peer assessment research. *The International Review of Research in Open and Distributed Learning*, *20*(5), 168–191. doi:10.19173/irrodl.v20i5.4333

Zheng, L. (2016). The Effectiveness of self-regulated learning scaffolds on academic performance in computer-based learning environments: A Meta-analysis. *Asia Pacific Education Review*, *17*(2), 187–202. doi:10.1007/s12564-016-9426-9

Wang, Z., Hwang, G.-J., Yin, Z., & Ma, Y. (2020). A Contribution-Oriented Self-Directed Mobile Learning Approach to Improving EFL Students' Vocabulary Retention and Second Language Motivation. *Educational Technology & Society*, *23* (1), 16–29.

A Contribution-Oriented Self-Directed Mobile Learning Ecology Approach to Improving EFL Students' Vocabulary Retention and Second Language Motivation

Zhuo Wang[1], Gwo-Jen Hwang[2], Zhaoyi Yin[3] and Yongjun Ma[1*]

[1]Qingdao University, China // [2]National Taiwan University of Science and Technology, China // [3]University of Glaskow, United Kingdom // zhuowang@qdu.edu.cn // gjhwang.academic@gmail.com // piperyin@126.com // qdmayongjun@163.com

[*]Corresponding author

ABSTRACT: Vocabulary mastery is critical to English as Foreign Language students. Mobile technologies enable students to learn vocabulary without space and time limitations. However, existing mobile-assisted vocabulary learning research often employed teacher-directed activities that increased instructors' workload, undermined student motivation or targeted individual cognitive outcomes only. In this study, a Contribution-oriented Self-Directed Mobile Learning Ecology (CSDMLE) model is proposed for developing student-directed and motivational vocabulary learning activities in groups. Through a mixed-method design, we administered a survey and a vocabulary test to 55 freshmen students in a Chinese university, and conducted follow-up interviews. We found that students in the CSDMLE group outperformed those not in the group in the post-test vocabulary test, but there was no statistically significant difference between the two groups' delayed vocabulary test or L2 motivation. However, the treatment group displayed a highly favorable attitude toward the learning approach and a strong intention to use it continuously. The findings have implications for technology-supported vocabulary learning activities.

Keywords: EFL, Mobile learning, Vocabulary retention, L2 motivation

1. Introduction

Vocabulary acquisition is fundamental in learning a second or foreign language (L2) (Hwang & Wang, 2016; Tight, 2010). Learners who master vocabulary well are more likely to produce better language performance. However, it is often a long and tedious process (Chen et al., 2019); if learners are not motivated to learn or they do not know how to learn effectively, they might give up learning L2 vocabulary (Dörnyei & Csizér, 2002). Generally, learners can acquire vocabulary effectively under instructors' guidance, but class time is often limited, and language instructors cannot guide and monitor students out of class with traditional learning activities.

The advancements in computing technologies have dramatically changed the way we live as well as how languages are learned (Hung et al., 2018). Among others, the use of mobile technologies has grown rapidly worldwide (Sundberg & Cardoso, 2019), as evidenced by the rapidly increasing rate of device ownership, and the wider coverage of mobile-cellular networks in both developed and developing countries (Kaliisa et al., 2019; Huang at al., 2010). Due to its advantageous affordances such as connectivity, ubiquity and interactivity (Klopfer et al., 2012), many researchers have conducted studies on technology use during language learning processes, including vocabulary learning (Gürkan, 2019). Meta-analyses of mobile-assisted language learning studies in the last decade indicated that vocabulary outcomes were the most frequently researched variable (Hwang et al., 2019; Elaish, 2019).

Although such studies attest to the benefits of mobile-assisted vocabulary learning (MAVL), they also unintentionally conveyed two misleading messages: one, MAVL has to be teacher-directed or demands tremendous effort from instructors; two, whether more words were retained is the most important criterion for activity design or platform selection. In many of the reviewed studies, the content sent out to students was either originally created or appropriately tailored from existing resources, which would place heavy demands on the instructors, both cognitively and physically (e.g., Pirasteh & Mirzaeian, 2015). The workload of creating and frequently distributing course vocabulary content might deter many language instructors from incorporating mobile learning to its full potential. On the other hand, while we acknowledge that vocabulary memorization and retention is critical to language success, it should be neither examined solely, nor over-emphasized to such an extent that overshadows learner interest, ease of use, motivation or other affective and perceptive factors. When learners feel overwhelmed by using a technology, it is probably they will terminate its use once they have a choice. Thus, researchers need to balance between pursuing

cognitive outcomes and catering to students' attitudinal and emotional needs when they integrate MAVL approaches for long-term use.

In all, the above needs demand MAVL designs that not only alleviate instructor workload, but also examine cognitive and affective changes with equal attention. One of the most important affective factors in language learning is second language (L2) motivation. Distinguished from generic motivation, a term that is often loosely used to encompass various emotional aspects, L2 motivation refers specifically to one's motivation to acquire a second or foreign language (Dörnyei, 2005). It is measured with certain established instruments worldwide, such as AMTB (attitude and motivation test battery). Understanding whether and how MAVL could impact one's L2 motivation may provide insightful guidance for language instructors.

To address the above challenges in MAVL research, we propose a Contribution-oriented Self-Directed Mobile Learning Ecology (CSDMLE) model to comprehensively guide our design of effective MAVL experience. Specifically, the model was hypothesized to improve students' vocabulary retention better than traditional approaches through utilizing related pedagogical theories (i.e., theory of multimedia learning); to reduce instructor workload, we asked students to create and share vocabulary learning content themselves; to boost their L2 motivation, we set goals that stimulated their sense of contribution and responsibility. Accordingly, our research questions are listed as follows:

- Is there any significant difference in the post and delayed vocabulary retention performance of those who learned via the CSDMLE model and those who did not?
- Is there any significant difference in the L2 motivation of the participants using the CSDMLE model and those using conventional learning?
- How do students in the CSDMLE group perceive this learning approach?

2. Literature review

2.1. Vocabulary acquisition and retention

Vocabulary plays an indisputably vital role in students' L2 learning (Tight, 2010). Historically, language learners have expressed significant difficulty remembering vocabulary words (Chen & Chung, 2008), and retention is one of the most difficult learning problems to address, due to the unavoidable forgetting nature of human beings (Ebbihaus, 1913). Research on cognitive science has suggested that a list of principles be followed to enhance long-term memory of learning materials, such as using images or graphics to assist verbal learning (Driscoll, 2005). Regardless, traditional vocabulary instruction is often limited in terms of both class time invested and effective retention strategies employed. Students need to mainly rely on themselves for vocabulary learning, which could create problems and frustration for student learners and result in their loss of motivation (Dörnyei & Csizér, 2002).

Two prominent approaches were often used to elicit successful vocabulary learning and retention: multimodal presentation and spaced repetition (Kohnke et al., 2019). The former entails supporting word understanding and retrieval with multimedia, which is underpinned by the Cognitive Theory of Multimedia Learning (Mayer, 2009). It postulates that word knowledge is acquired through visual and verbal channels. When learners obtain word knowledge with multiple media forms, it stimulates both channels and strengthens one's memory retention. Spaced repetition refers to a programmed system with designated time intervals that provides a series of presentations or practices of vocabulary content (Kohnke et al., 2019). Being regularly and rhythmically exposed to word knowledge, learners can efficiently maximize their understanding and elongate their knowledge retention (Pellicer-Sánchez & Schmitt, 2010). Effective use of both the multimodal and spaced repetition approaches promises to yield satisfactory vocabulary learning outcomes.

2.2. MAVL research trends and limitations

The integration of mobile technologies and devices in vocabulary learning has gradually led to the field of MAVL. There has been a steadily increasing number of MAVL studies since the last decade. More recently, several meta-analyses have been conducted to synthesize MAVL research trends and gaps on different levels, including effect size, research settings, aspect of vocabulary knowledge, study duration, etc. For example, Lin and Lin (2019) found

that learners generally displayed a positive attitude toward MAVL. Mahdi (2018) concluded that receptive knowledge was exploited more frequently than productive knowledge. In this paper, we aimed to highlight certain limitations or gaps that warrant imminent attention for MAVL activity and research design.

First of all, in terms of goal-setting, most studies sought to improve individual outcomes, such as vocabulary retention and learning interest. For example, Alemi et al. (2012) conducted research upon 45 freshmen students and found statistically significant difference in treatment and control groups' delayed test mean score. Chen et al. (2019) found that primary students who learned vocabulary via their app-based self-regulated mechanism improved vocabulary retention and motivation significantly better than those in control group. While these are indeed important learning objectives, they do not emphasize sense of community, collaboration or socialization skills that are in rapid need today.

Secondly, regarding the content design and form of MAVL activities, there is a lack of studies employing student-centered approaches, such as Constructivism or self-directed learning. Most existing MAVL research reported activities that were still teacher-directed that undermined learner autonomy. This reflects a Behaviorist epistemology, placing students as passive knowledge recipients (Hu, 2013). For instance, undergraduate students in Pirasteh and Mirzaeian's (2015) study were reported to receive phrasal verb content prepared by course instructors through SMS every day for 25 days. This not only limited student output or productive skill development, but also created additional workload for instructors. It is observed that teachers who integrate mobile learning often need to commit more effort, such as digitizing the content to be placed in mobile devices, ensuring functionality, and solving emergent technical difficulties (Shih et al., 2010). Even with positive results attained, it remains questionable if instructors who went through the tedious process of creating and distributing content would persist such an endeavor in a longer term. Few studies epitomized student-generated content and self-directed learning. For example, 24 Iranian EFL students in Foomani and Hedayati's (2016) study took photos to demonstrate word usage and shared them on Padlet for peer discussion, but the study employed a pre-experimental design that was mainly descriptive. Botero et al. (2019) examined whether using Duolingo out of class could promote 118 university language students' self-directed learning, and found that students lacked sustained motivation in such learning and needed stronger sense of responsibility. Wong and Looi (2011) reported two case studies in which primary students took photos and created sentences for class discussion, and advocated to treat student-generated content as the "end."

Thirdly, L2 motivation, which is the most reliable predictor of language learners' long-term effort in L2 learning (Dörnyei, 2005) was rarely examined in MAVL studies. Although the term "motivation" was often mentioned in MAVL studies, it has been used more as an umbrella term for constructs like learning interest, intention to use, satisfaction, and may denote meanings that vary from study to study. For example, in Looi's et al. (2011) research, motivation was depicted as students' attitude toward and engagement in mobile learning, and relevant results were obtained based on classroom observation and a self-designed survey. More recently, in Loewen et al.'s (2019) study, eight participants' motivation level was inferred from their learning journal, in which they described their interest in and mood for learning Turkish via Duolingo. In terms of L2 motivation, AMTB developed by Gardner (1985) is a widely used instrument among language scholars worldwide. For example, Jain and Sidhu (2013) in Malaysia used AMTB to measure freshmen students' L2 motivation, and found that increasing anxiety would reduce their level of motivation, regardless of discipline, gender or language proficiency. Rahmany et al. (2013) used AMTB to determine the L2 motivation level of 60 Iranian EFL of different age groups and found that extensive reading did not elicit better L2 motivation.

Meanwhile, although university students were frequent participants in MAVL studies (e.g., Yuan, 2019; Hanson & Brown, 2020), there is a lack of research on pre-service teachers. Yet, how their perception of English, and the way they were taught English could potentially impact their future teaching philosophy and performance to great extent. Thus, investigating how pre-service teachers might benefit from innovative learning interventions could have a far-reaching significance.

Overall, the various limitations identified above demand a more comprehensive framework that is grounded upon solid pedagogical and instructional theories, and provides clear guidance for MAVL design in terms of content to be used, form it takes and a goal that is motivating and yet practical. The following model was designed as a response to this demand.

3. The contribution-oriented self-directed mobile learning ecology model for vocabulary learning

To meet language learners' both cognitive and affective demands, and begin to address the identified gaps, we developed the Contribution-oriented Self-Directed Mobile Learning Ecology model, specifically informed by three theoretical frameworks. In this model (see Figure 1), the triangle represents the three pedagogical aspects that were identified in our literature review as lacking improvement, namely the goal, content and form of the MAVL design. Secondly, the inner circle consists of three corresponding patterns that are deemed as problematic. Thirdly, the outer circle depicts our CSDMLE model with three key components, which are in direct contrast with the previous approach displayed in the inner circle. For example, while the majority of MAVL studies focused on producing individual outcomes such as increased vocabulary test scores, our model advocates collaborative gains in addition to individual growth. Finally, each component in the outer circle is supported by and grounded upon a particular theory as introduced further below.

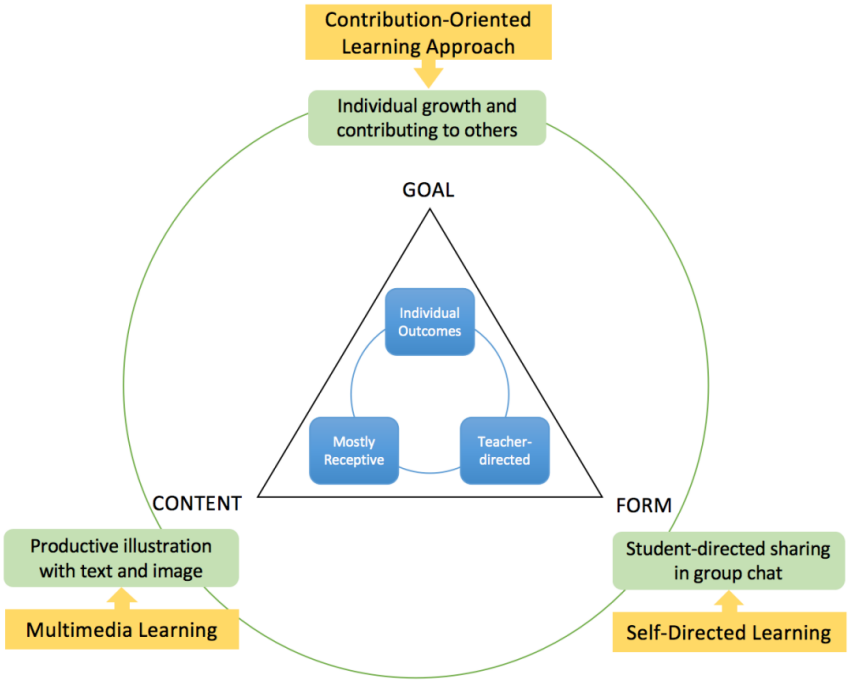


Figure 1. The contribution-oriented self-directed mobile learning ecology model

To begin with, the goal-setting was inspired by the Contribution-Oriented Learning Approach (COLA) proposed by Collis and Moonen (2001). COLA is a pedagogical theory that advocates the contributing role of individual learners in online environments. It characterizes the role of the instructor as facilitator and coordinator of activities, and that of students as learning resource creators and designers who should "contribute to make a difference" (Collis & Moonen, 2006). A distinct feature of COLA-informed activities is that students produce meaningful resources that can be practically used or reused by others for authentic purposes at a later time (Collis & Moonen, 2001). Such resources are in sharp contrast with traditional assignments that are often deemed as learning evidence and offer limited value beyond the individual students. The goal of making an actual contribution is believed to encourage students to take responsibility for their own learning, foster a sense of community as well as build a collaborative culture.

Next, the theory of Multimedia Learning, which guides an effective design of multimodal information presentation, was used to guide students' productive content design. Since its central premise is that using both verbal and visual channels is more effective than using either alone for promoting understanding and retention (Mayer, 2009), students were asked to create illustrations with text (contextualized sentence-making for a chosen word) and images (a corresponding picture that echoes the text) (see Figure 2).

The third theory is Self-directed learning (SDL) proposed by Knowles (1975). SDL is a "basic human competence--the ability to learn on one's own" (Knowles, 1975, p. 17), and it has been well researched in the field of adult education. Self-direction is perceived as a significant component of achieving meaningful educational outcomes (Garrison, 1997). In order to explain what SDL encapsulates, Garrison (1997) proposed a comprehensive model consisting of three fundamental yet highly interconnected dimensions, including self-management, self-monitoring and motivation. In particular, self-management refers to learners' active control during the learning process, but the control "must balance educational norms and standards with student choice and the responsibility for constructing personal meaning" (Garrison, 1997, p. 23). Self-monitoring encompasses the cognitive and meta-cognitive processes and refers to learners taking responsibility for active meaning making and critical reflection. Finally, motivational factors have a pervasive influence on learners' goal-setting and subsequent task effort (Corno, 1989). More importantly, the entering motivation or the motivation to enter into a task plays a significant role in learners' assessment of task value and attainability.

Figure 2. Two samples of student-created vocabulary illustrations

Congruently informed by all theories and components in the CSDMLE model, the activity was expected to proceed according to the following stages:

(1) Entering motivation: the researchers describe the activity and allow students to determine whether they want to participate after assessing the task difficulty on their own.
(2) Acquisition: students select a word from the required textbook glossary and study its meaning and usage.
(3) Self-managing: students actively control their learning pace, the resources they want to consult, the applications (apps) they want to use, and the extent to which they conform with the task standards. Specifically, each student should produce an illustration that displays both the chosen word's contextualized usage and an image that complements the text.
(4) Contribution: students post their illustrations to a designated group chat in WeChat, the most popular social media app in Mainland China, so that students in the same group can view and learn about the shared resources.
(5) Self-monitoring: Through viewing illustrations shared by other students in the group chat, learners actively compare and connect their own understanding and others' presentations, and reflect on the quality of and strategies used for their last illustration.
(6) Enhancing motivation: Feeling surprised or benefiting from others' illustrations, students are motivated to continually improve their own and produce quality content for peers.

4. Methodology

4.1. Research design

According to Creswell (2009), a sequential explanatory mixed-method design refers to using quantitative data collection first and qualitative methods later that builds on the former. Such a strategy is appropriate when researchers intend to explain and interpret quantitative results by collecting and analyzing follow-up qualitative data (Creswell, 2009), and can be especially helpful when unexpected findings emerge from a quantitative study (Morse, 1991). The purpose of the present study was to determine the effect of the CSDMLE model on students' vocabulary

retention and L2 motivation. Given the novelty of our model, we also anticipated that certain unexpected results might arise. Thus, follow-up qualitative data collection was added to help us interpret any perplexing findings.

4.2. Participants

Participants were two freshmen classes majoring in English at a Chinese normal university who were recruited via convenience sampling. Although the two groups were taught by different instructors of the same course, "Contemporary English," their college entrance examination scores did not differ statistically significantly. A total of 28 students were in the experimental group (25 females and 3 males) and there were 27 in the control group (24 females and 3 males). The mean ages were 18.89 and 18.96, respectively.

4.3. Procedure

Both classes were first invited to complete the consent form and a pre-study survey online. Upon receipt of their responses, the researcher randomly assigned the participating classes to either the experimental group (EG) or the control group (CG). Each group then created a group chat on WeChat.

4.3.1. Pre-training

Before the study began, the researchers first joined both group chats on WeChat, and disseminated the study requirements via PowerPoint slides for the respective groups. Students in either group were then given one day to raise any questions or concerns about the study. For the CG, researchers explained the dates and form of upcoming vocabulary tests and surveys, as well as the use of the group chat for such purposes; for the EG, beside tests and surveys, the researchers also described the steps taken to create a quality illustration, demonstrated exemplary illustrations to help EG students visualize what was required, and elaborated on the posting schedule. The instructors for either group were not only asked to undertake the same instructional practice in class (selecting the same key words and phrases to instruct directly, and spend the same amount of time on vocabulary instruction), but also invited into the group chat so that they were aware of all activities and could respond promptly if unexpected problems arose.

4.3.2. Study participation

As shown in Figure 3, after completing a survey and a subsequent vocabulary test online, the CG continued to learn vocabulary in class and used their conventional approaches at will; the EG learned vocabulary in the same way as the CG when in class, but created and shared illustrations in the group chat in their spare time out of class. The EG's schedule was as follows: posting their first illustration by 8pm on Tuesday, and their second by 8pm on Friday. The study lasted for 2 weeks, with two illustrations per student each week. Informed by Spaced Repetition, such a schedule spaced out the students' illustrations across the week, and increased their times of exposure to vocabulary knowledge, which would highly probably lead to more effective and efficient vocabulary retention. It should be noted that while the instructor was present in the EG's group chat, she had been politely asked to only intervene when an illustration contains incorrect information that was not timely revised by the author student him- or her- self. Immediately after the study ended, all students completed a survey and took the vocabulary test again. Two weeks later, all students took a delayed vocabulary test. Three weeks later, four EG students participated in the interview.

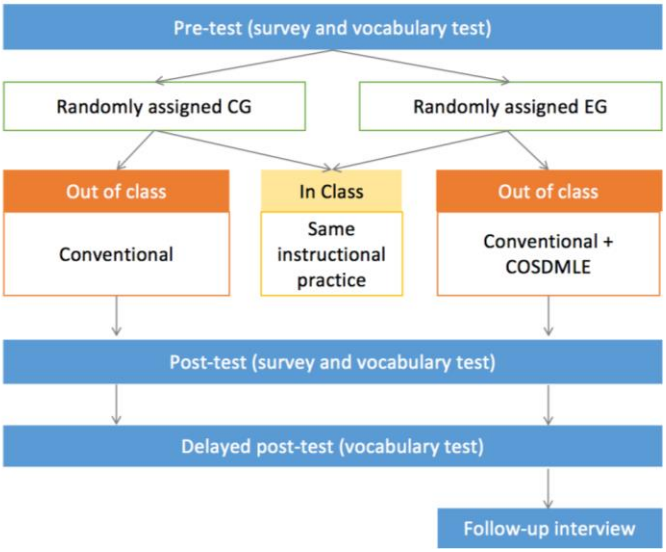


Figure 3. Study procedure

4.4. Instruments

The instruments used in the present study include a vocabulary test, a survey on L2 motivation, a questionnaire on learning experience and satisfaction, and an interview.

The vocabulary test took the form of active recall, asking students to provide the Chinese meaning for the given English lexical items. A total of 60 vocabulary items were selected by the English instructors as worth being included (considered new or difficult to memorize) in the target learning module. In random orders, these same items were tested in a pre-, post- and delayed-test fashion. Two raters scored the tests independently first, and then discussed those with differing opinions until they achieved agreement. The final analysis of data only included items that were illustrated by the students.

The L2 motivation survey was adapted from Gardner's (1985) AMTB, an instrument developed specifically to evaluate learners' L2 motivation. The Cronbach's α value is .90. In our study, three sub-scales were included: Attitude toward English Learning (ATEL), Motivational Intensity (MI), and Desire to Learn English (DTLE). ATEL evaluated respondents' general attitude toward English learning; MI assesses the intensity of a student's motivation to learn an L2 such as their effort in classroom assignments, future plans about their language study, etc. DTLE inquired about students' desire to learn an L2. A Chinese version of the survey, which has been validated by four English teachers through back-translation, was provided to the students.

The learning experience and satisfaction questionnaire has 18 items. It was developed by the researchers and then validated by two experts in the Instructional Technology field with over 10 years' experience. It asked about the EG's perceptions of their learning, usage of the illustrations and related behaviors. There were 13 multiple-choice questions, three checklists, one matrix, one ranking and one open-ended question.

Additionally, semi-structured interviews (through phone and instant messaging) were conducted with four EG participants to understand their perceptions and opinions based on the following questions:

- What do you consider the most beneficial features in this MAVL experience?
- What factors have hindered you from making the most of this learning?
- What did you do with the illustrations shared by others?
- What was the role of your English instructor?

5. Results

5.1. Vocabulary retention

The descriptive statistics for the three tests of EG and CG were presented in Table 1. We conducted an independent *t*-tests for pre-test, post-test and delayed test respectively (Table 2). Because of the voluntary nature of the study, some students chose not to participate in all three tests, and few missed the post-test or delayed test, during which they needed to address other priorities, such as course assignments or interest community meetings. Thus, the sample size varied in each test. The results indicated that there was statistically significant difference between the two groups' pre-test and post vocabulary test scores, but no statistically significance in the delayed test.

Specifically, in the pre-test, CG scored statistically significantly better than the EG (t = -2.60, p < .05). A possible explanation would be that some Chinese students had a habit of previewing or self-teaching learning content in upcoming modules in advance, so that they would understand better in class; therefore, these students would attain better scores even when being tested on content that was not taught yet. According to the results and line graph (Figure 4), EG caught up by the post-test and outperformed CG with statistical significance (t = 2.42, p < .05).

Table 1. Descriptive summary for vocabulary scores

	Group	*N*	Mean	Std. deviation	Std. error mean
Pre	EG	28	20.11	8.04	1.52
	CG	27	29.00	16.14	3.11
Post	EG	22	52.64	6.45	1.37
	CG	20	45.30	12.55	2.81
Delayed	EG	26	46.96	10.58	2.07
	CG	23	42.70	13.73	2.86

Table 2. Independent *t*-test results for vocabulary scores

	Levene's test for equality of variances		*t*-test for equality of means						
	F	*Sig.*	*t*	*df*	*Sig.*	Mean difference	Std. error difference	95% CI	
								Lower	Upper
Pre	13.45	0.00	-2.60	53.00	0.01	-8.89	3.42	-15.75	-2.03
Post	3.64	0.06	2.42	40.00	0.02	7.34	3.04	1.20	13.48
Delayed	0.42	0.52	1.23	47.00	0.23	4.27	3.48	-2.73	11.27

Figure 4. Vocabulary test mean score plot

5.2. L2 motivation

According to the two-way mixed ANOVA (see Table 3), there was no statistically significant interaction between treatment and time on motivation, F (1, 51) = 0.51, $p > .05$, partial $\eta^2 = 0.01$. In other words, the two groups were not statistically significantly different in terms of L2 motivation at pre- or post-test. Visually, it can be seen in Figure 5 that EG increased slightly more than the CG from pre-test to post-test, but the result needs to be interpreted with data from the survey and interview holistically.

Table 3. Two-way mixed ANOVA summary table for motivation

Source	Type III sum of squares	*df*	Mean square	*F*	*Sig.*	Partial eta squared
Time	0.08	1	0.08	1.81	.18	0.03
Time * Group	0.02	1	0.02	0.51	.48	0.01
Error(Time)	2.28	51	.05			

Table 4. Correlation analysis for design factors predicting post L2 motivation

	Assistance of image	Sense of rapport	Sense of contribution	Dictionary use
Post-L2Motiv	0.526**	0.514*	0.579**	0.571*

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

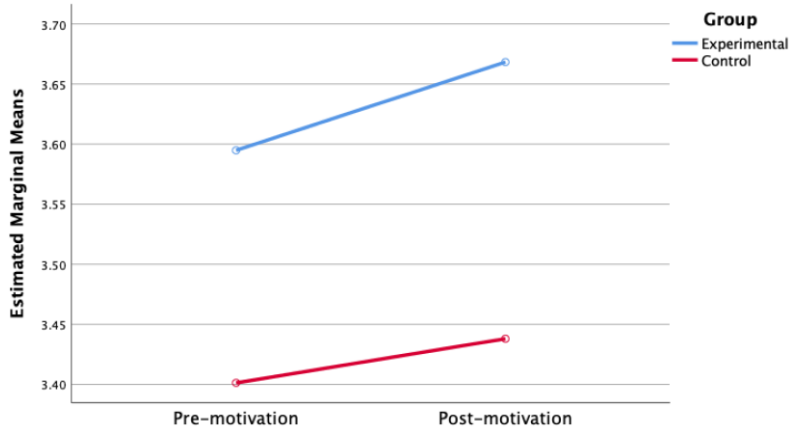


Figure 5. L2 motivation graph

A correlation analysis between students' perceived effective features of the treatment and the L2 motivation results was conducted to see what exact features predicted the EG's post L2 motivation, and the results are shown in Table 4. It can be seen that all the listed factors were statistically significantly correlated with post L2 motivation, with "sense of contribution" the strongest, followed by "dictionary use," "assistance of image" and "sense of rapport." Simply put, participants who acknowledged more of the image incorporation and propelled dictionary use, and who felt a stronger sense of rapport and contribution were more motivated to learn English subject as a whole.

5.3. Learning perceptions and satisfaction

When asked whether they wished to continue to learn this way, 92.3% of the EG participants responded positively. When asked to compare with traditional learning approaches on various dimensions by responding more, neutral or less, EG rated our approach as more satisfactory (80.8%), easier to use (80.8%), more memorable (69.2%), more flexible (65.4%), more interesting (65.4%), and more efficient (53.8%).

A correlation analysis further suggested that factors listed in Table 5 were statistically significantly correlated with each other. For example, students who considered our approach more flexible than traditional learning approach were very likely those who also rated high memorability (r = .916, $p < .01$). Additionally, among other factors, student who found MAVL more interesting were most likely to yield a higher level of satisfaction overall.

Table 5. Correlation analysis between perceived qualities

	Interestingness	Efficiency	Ease of use	Memorability	Flexibility
Efficiency	.636**				
Ease of use	.671**	.427*			
Memorability	.566**	.729**	.732**		
Flexibility	.660**	.778**	.671**	.916**	
Satisfaction	.671**	.597**	.505**	.520**	.465*

Note. $^{*}p < .05$; $^{**}p < .01$.

5.4. Interview results

In order to gain an in-depth understanding of EG students' experiences, perceptions and motivation, we also conducted follow-up interviews. Though we attempted to recruit six EG participants with various level of commitment to MAVL, only four participants who consented to the interview were actually interviewed (one highly committed, three moderately). One was conducted over the phone, and the other via WeChat text messages.

Regarding the strengths of MAVL, all four participants mentioned that learning was more flexible and personalized this way. Three out of four stated that the illustrations created by their classmates were very helpful and of high quality. One student mentioned that such illustrations were "very down-to-earth, and conveyed a sense of proximity...that those standardized ones in the textbook or found online would never be able to achieve." Another student noted that MAVL helped build collegiality among classmates, because "whenever someone posts in the group chat, it was like saying, I'm studying vocabulary now or I'm with you." The frequent posted illustrations were also seen as automatic reminders to study or create illustrations, as the more often students reviewed the content, the better they would retain such knowledge.

In terms of ideal changes, one important note was that the retrieval of shared illustrations was a little cumbersome, and they wished to have someone store and organize them in a public folder every week. The number of required illustrations was mentioned as well, with some indicating more would be better and some preferring to create just one, because "the instructor may create a bigger group chat and invite students from other classes, so that we reap more but contribute less individually." It was also pointed out that sometimes students would illustrate the same word, and not bother creating another one, which could leave out some advanced or difficult words.

With regards to learning behaviors, all participants said that they would use a dictionary or discuss with friends when they were unsure about the usage of words illustrated by their classmates, especially when a word has multiple meanings or properties. It was emphasized by one student that normally she would seldom use a dictionary, because the textbook glossary was sufficient for understanding a word's meaning, but in order to be 100% sure about her illustrations, she had to use the dictionary more often, so that other students could learn correct knowledge. This was consistent with the survey data for Question 16, which asked students about their perceptions of their own illustrations: 73% of the respondents chose the option "I only sent out illustrations when I was 100% sure about the image and text accuracy." This indicated that students felt accountable for information they shared, and were thus more cautious of potential errors in vocabulary use than when they learned individually. Students also mentioned that they often guessed the meaning before using the dictionary if a word was unfamiliar, because "it was more fun...you can gain more confidence if you guessed right." One participant said that she did not learn about the illustrations as soon as she saw them in the group chat; rather she saved all illustrations to her photo album, and viewed them when she had longer chunks of time.

As for the role of the instructor, the interviewees agreed that the presence of instructors in the group chat was necessary, because "it means the instructor considers the task important" or "it makes students feel more secure because he or she can help when we need it." At the same time, they acknowledged that how much the instructor should be involved was a challenging decision to make: "...if they are involved too much, it would be overstepping; if it is too little, we may not treat it seriously." One student added that "the bottom line is, the instructor should be encouraging rather than judgmental. It's helpful to let us know when we did something wrong, but it could also be discouraging or devastating to sensitive classmates, because this is a public space."

6. Discussion and conclusions

Both vocabulary retention and L2 motivation are crucial for long-term language success. In this paper, we attempted to examine how and which MAVL features can enhance both, so that instructors and researchers can make more informed decisions when adopting and developing such activities.

First, we found that EG students, who attained a statistically significantly lower mean vocabulary score at pre-test than CG, outperformed in the immediate post-test with statistically significantly difference; however, there was no statistically significant difference in the groups' delayed test. Such a finding is consistent with Zhang et al.'s (2011), in which two sophomore classes (one SMS group, one paper group) were compared, and statistically significant difference was only found in the two groups' post-test scores, but not the delayed test. This could be an indication of MAVL's apparent effectiveness in improving initial vocabulary acquisition, while traditional approaches like rote memorization can still make up for such disadvantage at a later time. This is especially true when tests use close-ended questions that simply require students to recall, not to produce. For example, students who provide correct meanings for the same word may differ in their ability to use it accurately and meaningfully in a sentence. Such differences are unlikely to be captured via receptive tests, and thus may account partially for the inability to detect statistically significant results. Researchers also suggested that when there were few words to learn, the advantage of one approach (i.e., mobile learning) over the other might be too subtle to detect (Lu, 2008; Derakhshan & Kaivanpanah, 2011).

Secondly, in terms of motivation, there was no statistically significant difference either within group or between groups. However, results from the EG's post-test questionnaire and interviews suggested that a motivational increase might be yet to come. On the one hand, EG were satisfied with and thought high of this treatment experience. For example, 24 out of 26 students in the EG responded that they wished to continue to learn this way, and they attributed the highest score to the treatment's ease of use as contributing to MAVL effectiveness. This is consistent with Huang et al.'s (2007) study on 313 undergraduate and graduate students' use of mobile learning, where they found a statistically significant correlation between one's perceived ease of use and their intention to use it. Additionally, interviewees acknowledged the flexibility, sense of collegiality and quality of peer-generated illustrations which all inspired them to continuously learn this way. On the other hand, our correlation analysis indicated that certain MAVL features (i.e., sense of rapport and contribution, and dictionary use) could strongly predict students' post-test L2 motivation. Thus, it may be expected that if students continue to gain benefits from these features, their L2 motivation will eventually increase after a longer period of use.

Thirdly, informed by the CSDMLE model, we integrated features that targeted effective goal-setting, content design and form adoption. Results from the post-test questionnaire, interviews, and group chat behavior observation showed that all three aspects were relatively successful. For example, we aimed to promote both individual growth and collaborative outcomes through contribution-oriented learning, and it was indeed found that students' sense of contribution was most statistically significantly correlated with their L2 motivation; individuals attained higher scores from pre-test to post-test, and the few illustrations created by each student aggregated to a larger collection of high quality learning materials, which was a testament to their collaboration and contribution. This is consistent with Alghamdy's (2019) finding that students enjoyed sharing with others in the mobile language learning environment. In terms of content design, most students could meet the activity requirements and created illustrations that contained both sentence(s) and an image. Results showed that the use of image was correlated to their post-test L2 motivation, meaning image incorporation was an valuable feature in MAVL. Consistent with SDL, we expected students to follow the MAVL prerequisites and timely create and share illustrations during the study. According to the questionnaire, 94.5% of the EG respondents claimed that they met the requirements well. Simply put, students were able to autonomously persist in this activity with little instructor interference.

The interview results showed that students generally appreciated the benefits of the CSDMLE-informed MAVL design, including its flexibility and repeated encounters with vocabulary that were often reported in other MAVL studies (e.g., Liu, 2016). Congruent with COLA and ML, our participants deemed helping others and the use of both image and text as essential for deep learning. Also, consistent with the quantitative results, students appreciated the sense of rapport and making contribution. Moreover, they suggested that instructors encourage instead of judging in such activities. This not only echoed Knowles's (1985) SDL theory, emphasizing learners' active control of the learning process, but also partially supported Chien's et al. (2020) finding that teachers' criticism might harm EFL

students' performance and confidence. Additionally, it was pointed out that the difficulty of message retrieval and lack of coverage of important vocabulary needed improvement.

Overall, the CSDMLE model was effective in guiding student-directed collaborative MAVL design. The mixed results from vocabulary test, L2 motivation, questionnaire and interviews suggest that students' satisfaction with, and inclination to participate in, MAVL is impacted by multifaceted factors. However, vocabulary retention, which is often stressed by most language instructors and researchers, did not seem as much an important concern to students in this research. The questionnaire analysis showed that whether students were satisfied with MAVL was statistically significantly predicted by the perceived interestingness, efficiency, sense of rapport, ease of use, memorability and flexibility. Indeed, ease of use has always been identified as a critical indicator of users' intention to adopt a technology (Lee, Cheung & Chen, 2005), and engaging factors are also valued by MAVL students (Attewell & Webster, 2005). Therefore, instructors who aim to adopt MAVL should design learning experience that promotes these aspects.

Finally, there is still room for improvement. For example, more participants or a multiple-stage design could have increased the finding's generalizability; the study may also have been carried out for a longer duration so that students' L2 motivation change could be more observable. Moreover, owing to the voluntary nature of the present study, some EG students did not commit fully to this learning experience or take the tests seriously, which may have discounted their own and peers' test performances. Additionally, vocabulary performance may need to be measured in more innovative and diverse forms, so that students' progress can be accurately captured. It is also advised to use multiple instruments, including both vocabulary tests and those that evaluate their affective changes which are either conducive to or the result of students' cognitive growth. Lastly, the study may incorporate design elements that distinguish between high-, intermediate- and low-proficiency students, so that different groups can benefit the most.

Acknowledgement

We especially thank Dr. Sarah Lohnes Watulak and Dr. Scot McNary for providing valuable suggestions for this study.

References

Alemi, M., Sarab, M. R. A., & Lari, Z. (2012). Successful learning of academic word list via MALL: Mobile assisted language learning. *International Education Studies, 5*(6), 99-109. doi:10.5539/ies.v5n6p99

Alghamdy, R. Z. (2019). The Impact of mobile language learning (WhatsApp) on EFL context: Outcomes and perceptions. *International Journal of English Linguistics, 9*(2), 128-135. doi:10.5539/ijel.v9n2p128

Attewell, J., & Webster, T. (2005). Engaging and supporting mobile learners. *Mlearn, 2004*, 15–19.

Botero, G., Questier, F., & Zhu, C. (2019). Self-directed language learning in a mobile-assisted, out-of-class context: Do students walk the talk? *Computer Assisted Language Learning, 32*(1–2), 71–97. doi:10.1080/09588221.2018.1485707

Chen, C. M., Chen, L. C., & Yang, S. M. (2019). An English vocabulary learning app with self-regulated learning mechanism to improve learning performance and motivation. *Computer Assisted Language Learning*, *32*(3), 237–260. doi:10.1080/09588221.2018.1485708

Chen, C. M., & Chung, C. J. (2008). Personalized mobile English vocabulary learning system based on item response theory and learning memory cycle. *Computers & Education, 51*(2), 624–645. doi:10.1016/j.compedu.2007.06.011

Chien, S. Y., Hwang, G. J., & Jong, M. S. Y. (2020). Effects of peer assessment within the context of spherical video-based virtual reality on EFL students' English-speaking performance and learning perceptions. *Computers & Education,146*, 103751. doi:10.1016/j.compedu.2019.103751

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research, 53*(4), 445–459.

Collis, B., & Moonen, J. (2001). *Flexible learning in a digital world: Experiences and expectations*. London, UK: Kogan Page.

Collis, B., & Moonen, J. (2006). The Contributing student: Learners as co-developers of learning resources for reuse in web environments. In D. Hung, & M. S. Khine, (Eds.), *Engaged Learning with Emerging Technologies* (pp. 49–67). Springer. doi:10.1007/1-4020-3669-8_3

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oak, CA: Sage Publications, Inc.

Derakhshan, A., & Kaivanpanah, S. (2011). The Impact of text-messaging on EFL freshmen's vocabulary learning. *The Eurocall Review, 19*(19), 39–47.

Driscoll, M. P. (2005). *Psychology of learning and instruction* (3rd ed.). Boston, MA: Pearson Education, Inc.

Dörnyei, Z. (2005). *The Psychology of the language learner: Individual differences in second language acquisition*. Mahwah, NJ: Lawrence Erlbaum.

Dörnyei, Z., & Csizér, K. (2002). Some dynamics of language attitudes and motivation: Results of a longitudinal nationwide survey. *Applied Linguistics, 23*, 421–426.

Ebbinghaus, H., (1913). *Memory*. New York, NY: Teachers College, Columbia University.

Elaish, M. M., Shuib, L., Ghani, N. A., & Yadegaridehkordi, E. (2019). Mobile English Language Learning (MELL): A Literature review. *Educational Review, 71*(2), 257–276. doi:10.1080/00131911.2017.1382445

Foomani, E. M., & Hedayati, M. (2016). A Seamless learning design for mobile assisted language learning: An Iranian context. *English Language Teaching, 9*(5), 206. doi:10.5539/elt.v9n5p206

Gardner, R. C. (1985). *Social psychology and second language learning: The Role of attitudes and motivation*. London, UK: Edward Arnold.

Gardner, R. C. (2006). The Socio-educational model of second language acquisition: A Research paradigm. *EUROSLA Yearbook, 6*(2006), 237–260. doi:10.1075/eurosla.6.14gar

Garrison, D. R. (1997). Self-directed learning: Towards a comprehensive model. *Adult Education Quarterly, 48*, 18-33.

Gürkan, S. (2019). Effect of annotation preferences of the EFL students' on their level of vocabulary recall and retention. *Journal of Educational Computing Research, 57*(6), 1436-1467. doi:10.1177/0735633118796843

Hanson, A. E., & Brown, C. M. (2020). Enhancing L2 learning through a mobile assisted spaced-repetition tool: An Effective but bitter pill? *Computer Assisted Language Learning, 33*(1–2), 133–155. doi:10.1080/09588221.2018.1552975

Hu, Z. (2013). Emerging vocabulary learning: From a perspective of activities facilitated by mobile devices. *English Language Teaching, 6*(5), 44–54. doi:10.5539/elt.v6n5p44

Huang, J., Lin, Y., & Chuang, S. (2007). Elucidating user behavior of mobile learning: A Perspective of the extended technology acceptance model. *The Electronic Library, 25*(5), 585–598. doi:10.1108/02640470710829569

Huang, A. F. M., Yang, S. J. H., Hwang, G. J., & Tsai, C. C. (2010). Situational language teaching in ubiquitous learning environments. *Knowledge Management & E-Learning, 2*(3), 312-328.

Hung, H., T., Yang, J. C., Hwang, G. J., Chu, H. C., & Wang, C. C. (2018). A Scoping review of research on digital game-based language learning. *Computers & Education, 126*, 89-104.

Hwang, G. J., Chu, H. C., Shih, J. L., Huang, S. H., & Tsai, C. C. (2010) A Decision-tree-oriented guidance mechanism for conducting nature science observation activities in a context-aware ubiquitous learning environment. *Educational Technology & Society, 13*(2), 53–64.

Hwang, G. J., & Fu, Q. K. (2019). Trends in the research design and application of mobile language learning: A Review of 2007–2016 publications in selected SSCI journals. *Interactive Learning Environments*, *27*(4), 567–581. doi:10.1080/10494820.2018.1486861

Hwang, G. J., & Wu, P. H. (2014). Applications, impacts and trends of mobile technology-enhanced learning: A Review of 2008-2012 publications in selected SSCI journals. *International Journal of Mobile Learning and Organisation*, *8*(2), 83–95. doi:10.1504/IJMLO.2014.062346

Jain, Y., & Sidhu, G. K. (2013). Relationship between anxiety, attitude and motivation of tertiary students in learning English as a second language. *Procedia - Social and Behavioral Sciences, 90*(InCULT 2012), 114–123. doi:10.1016/j.sbspro.2013.07.072

Kaliisa, R., Palmer, E., & Miller, J. (2019). Mobile learning in higher education: A comparative analysis of developed and developing country contexts. *British Journal of Educational Technology, 50*(2), 546–561.

Klopfer, E., Sheldon, J., Perry, J., & Chen, V. H. H. (2012). Ubiquitous games for learning (UbiqGames): Weatherlings, a worked example. *Journal of Computer Assisted Learning, 28*, 465–476.

Knowles, M. S. (1975). *Self-directed learning: A guide for learners and teachers*. New York, NY: Association Press.

Kohnke, L., Zhang, R., & Zou, D. (2019). Using mobile vocabulary learning apps as aids to knowledge retention: Business vocabulary acquisition. *Journal of Asia TEFL*, *16*(2), 683–690. doi:10.18823/asiatefl.2019.16.2.16.683

Lee, M. K. O., Cheung, C. M. K., & Chen, C. Z. (2005). Acceptance of Internet-based learning medium: The Role of extrinsic and intrinsic motivation. *Information & Management, 42*(8), 1095-1104.

Lin, J. J., & Lin, H. (2019). Mobile-assisted ESL/EFL vocabulary learning: A Systematic review and meta-analysis. *Computer Assisted Language Learning, 32(*8), 878–919. doi:10.1080/09588221.2018.1541359

Liu, P. L. (2016). Mobile English vocabulary learning based on concept-mapping strategy. *Language Learning and Technology, 20*(3), 128–141.

Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F., & Rawal, H. (2019). Mobile-assisted language learning: A Duolingo case study. *ReCALL, 31,* 293–311. doi:10.1017/S0958344019000065

Looi, C. K., Zhang, B., Chen, W., Seow, P., Chia, G., Norris, C., & Soloway, E. (2011). 1:1 mobile inquiry learning experience for primary science students: A Study of learning effectiveness. *Journal of Computer Assisted Learning, 27*(3), 269–287. doi:10.1111/j.1365-2729.2010.00390.x

Lu, M. (2008). Effectiveness of vocabulary learning via mobile phone. *Journal of Computer Assisted Learning, 24*(6), 515-525.

Mahdi, H. S. (2017). The Use of keyword video captioning on vocabulary learning through mobile-assisted language learning. *International Journal of English Linguistics, 7*(4), 1-7. doi:10.5539/ijel.v7n4p1

Mayer, R. E. (2009). *Multimedia learning* (2nd ed). New York, NY: Cambridge University Press.

Motallebzadeh, K., & Ganjali, R. (2011). SMS: Tool for L2 vocabulary retention and reading comprehension ability. *Journal of Language Teaching and Research, 2*(5), 1111–1115. doi:10.4304/jltr.2.5.1111-1115

Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do things fall apart? *Reading in a Foreign Language*, *22*(1), 31–55.

Pirasteh, P., & Mirzaeian, V. (2015). The Effect of short message service (SMS) on learning phrasal verbs by Iranian EFL learners. *Language in India*, *15*(1), 144–161.

Rahmany, R., Zarei, A. A., & Gilak, S. (2013). The Effect of extensive reading on Iranian EFL learners' motivation for speaking. *Journal of Language Teaching and Research, 4*(6), 1238–1246. doi:10.4304/jltr.4.6.1238-1246

Shih, J.-L., Chuang, C.-W., & Hwang, G.-J. (2010). An Inquiry-based mobile learning approach to enhancing social science learning effectiveness. *Educational Technology & Society, 13* (4), 50–62.

Stockwell, G. (2007). Vocabulary on the move: Investigating an intelligent mobile phone-based vocabulary tutor. *Computer Assisted Language Learning, 20*(4), 365–383. doi:10.1080/09588220701745817

Sundberg, R., & Cardoso, W. (2019). Learning French through music: The Development of the Bande à Part app. *Computer Assisted Language Learning*, *32*(1–2), 49–70. doi:10.1080/09588221.2018.1472616

Suwantarathip, O., & Orawiwatnakul, W. (2015). Using mobile-assisted exercises to support students' vocabulary skill development. *TOJET: The Turkish Online Journal of Educational Technology*, *14*(1), 163–171.

Tight, D. G. (2010). Perceptual learning style matching and L2 vocabulary acquisition. *Language Learning*, *60*(4), 792–833.

Wong, L.-H., & Looi, C.-K. (2010). Vocabulary learning by mobile-assisted authentic content creation and social meaning-making: Two case studies. *Journal of Computer Assisted Learning, 26*, 421-433.

Yuan, Y. (2019). Empirical study on the mobile app-aided college English vocabulary teaching. *International Journal of Engineering and Technology, 11*(1), 68–74. doi:10.7763/ijet.2019.v11.1125

Zhang, H., Song, W., & Burston, J. (2011). Reexamining the effectiveness of vocabulary learning via mobile phones. *Turkish Online Journal of Educational Technology-TOJET, 10*(3), 203–214.

Facilitating Communicative Ability of EFL Learners via High-Immersion Virtual Reality

Fang-Chuan Ou Yang[1], Fang-Ying Riva Lo[2], Jun Chen Hsieh[3] and Wen-Chi Vivian Wu[3,4*]
[1]Computer Science & Communication Engineering, Providence University, Taiwan // [2]Center for General Education, Asia University, Taiwan // [3]Department of Foreign Languages and Literature, Asia University, Taiwan // [4]Department of Medical Research, China Medical University Hospital, China Medical University, Taiwan // ouyang18315@pu.edu.tw // flo@asia.edu.tw // curtis3883@asia.edu.tw // vivwu123@asia.edu.tw
[*]Corresponding author

ABSTRACT: Developing communicative ability of English as a Foreign Language (EFL) learners is essential when it comes to authentic learning. Nevertheless, conventional textbook usage and English instruction often fail to be learner-engaging. With the help of high-immersion Virtual Reality (VR), language learning can be transformed into a more self-directed learning experience, using a simulated authentic environment to enhance engagement. Therefore, a three-dimensional learning system, Virtual Reality Life English (VRLE), was developed to provide learners with an authentic setting to facilitate communicative ability development. Seventy-two low-achieving junior high school students were recruited as participants. Multiple data sources were collected for both quantitative and qualitative data analysis of VRLE, including a pre-test/post-test addressing communicative performance, an Igroup Presence Questionnaire (IPQ) for the students' perception of perceived presence, and a semi-structured interview. The primary affordances were the beneficial application of VRLE to English communicative ability and an enhanced sense of presence in an EFL context. Furthermore, the students were positive about the learning experience. The study proves the potential of incorporating high-immersion VR technology in an EFL context. Nevertheless, the challenge of its accessibility needs careful consideration in future research to place VR in an advantageous position for language learning.

Keywords: Virtual reality, Presence, Immersion, Communicative ability, English as a Foreign Language

1. Introduction

Textbooks and related learning materials continually evolve to make studying English as a Foreign Language (EFL) more diverse and less challenging, often incorporating multimedia learning materials, such as video/audio CDs or MP3 files (Wang, Lin, & Lee, 2011). However, these materials often fall short of the ever-changing needs by providing a static and conventional paradigm, rather than an interactive representation of language (Lee & Chen Hsieh, 2018; Matsuda, 2017; McKay & Brown, 2016). In addition, most textbooks offer very limited opportunities for learners to engage in an authentic, meaningful learning context. Students are thus faced with the lack of practical contexts in the learning process (Chien, Hwang, & Jong, 2020). English is, therefore, often regarded as a traditional subject to master in the classroom rather than as a living language to be developed for exploring the real world (Chen Hsieh, Wu, & Marek, 2017), which is especially true in the EFL context.

Actual communicative ability, however, is crucial in English learning (Canale, 2014). Learning to communicate has been regarded as one of the greatest obstacles for EFL students (Zhang & Liu, 2018), since communicative ability not only encompasses inherent grammatical competence but also requires employing norms of usage and appropriateness in a variety of communicative situations (Hymes, 1972). Communicative ability is operationally defined in this study as the skillset and ability to achieve communicative goals in a contextualized setting, referring to the ability of EFL learners to employ English successfully in real-world situations, such as getting meaning across (see Abed, 2011; Rivers, 1972; Rouhi & Saeed-Akhtar, 2008) or responding to question prompts in a shop.

The rapid development of technology has made EFL learning and instruction more dynamic and has shifted the linguistic focus from grammar, vocabulary memorization, and sentence-mimicking to communicative applications. Online platforms, such as social networking sites (e.g., Barrot, 2016), Wikis (Zou, Wang, & Xing, 2016), and blogs (Pham & Usaha, 2016), have strengthened this new focus. While the importance of communicative skills in English learning is widely recognized by both instructors and learners, evaluation of English ability usually focuses on grammar and vocabulary as a detached part of the language. Whether learners can truly "use" English to solve problems and to communicate successfully in their actual daily lives is often neglected by instructors, especially in

junior high school. Since traditional English pedagogy is weaker in assisting EFL learners to achieve communicative objectives, there has been an urgent call for language instructors to adopt new techniques and tools to empower learners with the ability to get their meaning across by reacting and responding in a natural communication context (Luo, 2017).

Virtual Reality (VR) is gaining attention among language instructors because it transforms traditional learning materials into a live and self-directed interactive learning experience, thus increasing both motivation (Lanier, 2017) and language performance (Chen, 2016). VR allows learners to interact and immerse themselves in an authentic learning context without physically leaving the classroom (Huang, Rauch & Liaw, 2010). Other advantages of VR include providing experiential or contextualized learning, enabling learners to make meaningful connections, promoting active learning, boosting confidence and motivation, and fostering engagement (see Dawley & Dede, 2014; González-Lloret & Ortega, 2014; Sadler et al., 2013; Wang, Anstadt, Goldman, & Mary, 2014).

In view of the aforementioned benefits of adopting VR in learning, the researchers of this study self-developed a three-dimensional animation VR English learning system using head-mounted display, called "Virtual Reality Life English" (VRLE), where learners were able to study, practice, and apply English to achieve communicative tasks by engaging and immersing themselves in a real-life simulated context. The researchers were motivated to design such a learning system by a factual long-lasting classroom experience observed at a junior high school in a rural area in Taiwan, where most students receive relatively meagre resource of English learning, thus resulting in an overall phenomenon of low-achieving students with low motivation in English. Given the described situation, one of the initial goals of designing the VR learning system was to light the fire of language learning by bridging the gap of real-life language applications with the help of emerging technology. Once the goal was achieved, students would find their own way to drive their future learning. Accordingly, this study examined the effectiveness this VR learning system on English communicative ability and sense of presence among low-achieving junior high students in Taiwan. In addition, language performance and learner perceptions about the VRLE system were also explored. This research attempted to address the following research questions:

RQ1. To what extent did VRLE facilitate the communicative ability of EFL low-achieving learners?
RQ2. To what extent did VRLE affect learners' presence in the virtual environment?
RQ3. What were the learners' overall perceptions of the VRLE system?

This study is significant because, although previous research has shown the benefits of using VR, its effects on EFL learning have remained under-explored (Dolgunsöz, Yildirim, & Yildirim, 2018), let alone probing into the use of head-mounted displays among low-achieving learners in EFL contexts. Connecting communicative ability with VR in EFL settings has also been insufficiently examined. Even fewer attempts have been made among low-achieving learners, in comparison with the relatively richer literature focusing on experienced or moderately proficient learners (Levak & Son, 2017), thus making empirical evidence particularly scarce about how VR facilitates junior high school students with low English proficiency and motivation concerning their communicative ability and other learning related factors. This study aimed to self-develop a high-immersion VR system as an intriguing material in an EFL setting and further extended prior research by narrowing its focus on the effects among high-immersion VR, low-achieving EFL learners, their communicative ability, sense of presence, and learning perceptions.

2. Literature review

The assertion that VR could be conducive to learning achievement is widely supported by the theories of constructivist learning, contextualized learning, and immersive learning. Constructivist learning holds that learning occurs when learners construct new understanding by connecting new information with prior knowledge, experiences, and background (Vygotsky, 1978). Thus, a pedagogical design that leads to, for instance, real-life social interactions, embedded learning, self-directed learning, and student-centered learning can foster positive learning experiences and achievement (Huang, Rauch, & Liaw, 2010; Piaget, 1969; Vygotsky, 1978). VR-supported learning supports the constructivist instructional design (Lin & Lan, 2015). In line with constructivist learning, contextualized learning echoes the importance of supplying relevance between new information learned and existing knowledge. Students often are not shown the connection between their school learning and real-life applications (Hu-Au & Lee, 2017). Thus, learning framed with a context via VR empowers learners to visualize the purpose of learning in a more heuristic manner. Finally, immersive learning in language education allows learners to be naturally engaged in an

"embodied and perception-action rich context" (Legault et al., 2019, p.2). Immersion is also considered to be imperative to enhance communicative ability and language mastery (Wang, Petrina & Feng, 2017). As the immersive technology of VR advances, virtual immersion can be promising in boosting an authentic learning environment (Lin & Lan, 2015). Further, the levels and the types of immersion offered via VR have transformed over time, leading to a variety of VR systems in the market that suit various needs.

2.1. Low-immersion VR vs. high-immersion VR

VR has been broadly defined as a representation of an environment by simulation or replication where users can self-explore and interact (Lee & Wong, 2014; Makransky & Lilleholt, 2018). In a broad spectrum, VR can be divided into low-immersion and high-immersion, based upon how graphics are displayed (Checa & Bustillo, 2020; Kaplan-Rakowski & Gruber, 2019). Users conduct low-immersion VR, also named desktop VR, in a conventional computer-based environment, whereas high-immersion VR requires head-mounted displays (HMDs) or surrounding projection screens in a room setting (Estes, Dailey-Hebert, & Choi, 2016; Freina & Ott, 2015). While low-immersion VR is more accessible and cost-effective in educational applications, high-immersion VR delivers high interaction and a greater sense of immersion that cannot be replaced by low-immersion VR (Chang, Hsu, & Jong, 2020; Freina & Ott, 2015). As the technology of VR continues to evolve, learners have a better illusion and perception of being situated in the virtual world personally instead of experiencing it through an avatar. The VRLE adopted in this research was a high-immersion VR technology with real-person dubbing, motion capture, body-tracking interaction, and a VR controller to allow higher immersion to occur naturally, meanwhile, offering a near-authentic simulated context for language learning.

2.2. VR applications in language learning

VR use has proliferated in a wide array of disciplines, including language education, because it offers simulated scenarios that keep the engagement of users at a higher level. Chen (2016) noted that the virtual environment in the online platform Second Life could provide learners with visual and linguistic stimuli to facilitate language teaching and learning. In Lan's study (2015), the positive results confirmed that the usage of virtual contexts in EFL learning could: (1) provide students with learning opportunities without time and space limitations, (2) provide students with a game-like scenario for English learning, and (3) enhance the language performance of EFL learners. Another study conducted by Lan (2014) confirmed that VR enhances overall speaking skills and positive learning attitudes. Levak and Son's (2017) study affirmed that the listening comprehension of the learners was increased. Similar results in a study by Hassani, Nahvi, and Ahmadi (2016) suggested higher language proficiency and lower grammatical mistakes in the learners' performance through VR application. Further, communicative ability training can sometimes arouse public speaking anxiety. VR, however, offers learners a safe-to-fail environment and encourages trial-and-error learning (Chien, Hwang, & Jong, 2020; Chou, 2018). VR also enables natural interactions in an immersive simulated environment to enhance communicative ability that nearly no other type of media offers. The above studies also indicated that virtual reality could be a beneficial way to overcome the barrier of a limited EFL learning environment, supplying physically or psychological immersive situations for students to truly apply their English in response to communicative practices, hypothesized with a high sense of embodied presence in the virtual world (Vrellis, Avouris, & Mikropoulos, 2016).

2.3. Presence in VR

Presence, defined by Lee (2004) as the "psychological similarities between virtual and actual objects when people experience – perceive, manipulate, or interact with – virtual objects" (p. 38), is the key factor that shows the effectiveness of VR in various contexts. To be more specific, presence is the appeal of VR, creating the illusion for users that they are actually in the virtual world. From the perspective of language learning, presence in VR has the potential to immerse learners in the target culture, with which most EFL learners do not have frequent access. Witmer and Singer (1998) defined presence as the subjective experience of being in one place or environment, even when one is physically situated in another. However, presence is actually a complex and multidimensional perception that is generated through an interplay of multi-sensory information and various cognitive processes (Diemer et al., 2015). In this sense, presence is a normal awareness phenomenon that requires directed attention. It is

based in the interaction between sensory stimulation, environmental factors that encourage involvement and enable immersion, and internal tendencies to become involved. Wang, Petrina, and Feng (2017) said that the ultimate VR design incorporated in education should strive for both immersion and presence. The higher the perceived perception and awareness in VR, the higher immersion and engagement the user would experience in VR-assisted learning. Therefore, creating a strong presence in VR is one of the major goals in designing a VR language lesson, so that users become fully immersed in the language learning process. Presence, therefore, translates into higher motivation, which in turn translates into more confidence for using English, and in the long term higher EFL ability (Wu, Yen, & Marek, 2011).

3. Method

3.1. Participants

The participants in this study were 72 junior high school students in the ninth grade in central Taiwan. None had experience using VR systems before the study. The participants included 36 males and 36 females. Based on the long-accepted consensus in the local context of Taiwan, and since the school was located in a relatively marginal area in central Taiwan, the participants were in a disadvantaged learning environment due to limited English resources available to them. Their regular English instruction normally was limited to three hours a week, with the class time dominated by teachers lecturing on grammar and vocabulary. Only a small portion of class time was left for actual communicative experiences. According to their performance on the Comprehensive Assessment Program for Junior High School Students, around 80% of the participants were considered low-achieving learners. Their English proficiency fell between Al and A2 on the Common European Framework of Reference for Language (CEFR), indicating their ability to understand and use basic expressions related to areas of most immediate relevance or communicating on familiar and routine matters. Overall, despite formal English instruction received at school, these participants were regarded as low-achieving learners in English learning.

3.2. VRLE design

Rather than adopting commercial learning systems, the researchers developed a VR learning system, named VRLE, specifically for low-achieving students in a disadvantaged English learning environment, with a goal of increasing their motivation to learn English. To meet the students' needs, four experienced junior high school English teachers were consulted about the content of the system. Since the researchers designed VRLE as an alternative teaching material to the junior high school-based curriculum, the contents targeted commonly seen daily life conversations, including making reservations at hotels and restaurants, purchasing a toy, asking for directions, and ordering a meal. One of the learning scenarios, for example, was a toyshop where the students were required to purchase an assigned toy through communicating with the virtual clerk. Given the chosen scenario, the English instructors scripted scenes for the simulated contexts for life English learning, including English dialogues suitable for the low-achieving junior high school students. Then the system development team constructed the 3D models, multimedia contents, programs, and VR interactions according to the scripts. VRLE was developed based on Unity. The 3D model used 3Ds MAX and VR HMD using HTC VIVE and Kinect for image recognition and full-body interactive systems. Since this study aimed to create a learning environment where the students could practice authentic conversations in English, the VRLE system design incorporated the five factors that Usoh, Catena, Arman, and Slater (2000) identified as affecting the user perception of presence (as shown in Table 1).

Table 1. Comparison table of presence factors

Factors that affect presence	Elements in VRLE
High-resolution information display that enables participants to recognize the existence of the display devices	Realistic graphic design
Consistency of the displayed environment across all sensory modalities	Background sound effects Real-person dubbing Real-person motion capture

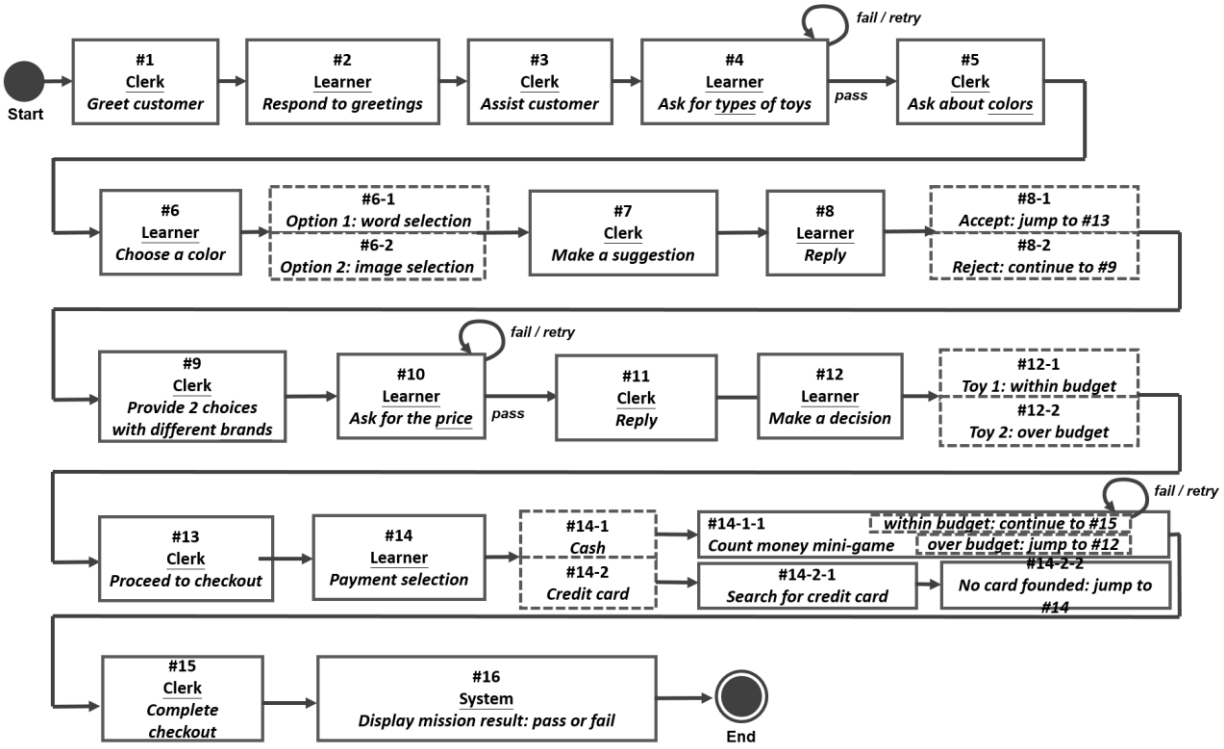
Being able to navigate through and interact with objects in the environment	Room scale setting Object taking
The virtual body should be similar in appearance or functionality to the individual's own body	Body tracking interaction
The connection between an individual's actions and effects of those actions should be simple	Flexible answering mode Pointing to select

3.2.1. The technical components of the system

- Realistic graphics: The settings in VRLE referenced real shop design and the characters in the system had body figures with real proportions.
- Background sound effects: Upon starting the learning task, users experienced realistic background sound effects that supported the visual surroundings, such as music or broadcasts commonly heard in a shop.
- Real-person dubbing: A native English-speaking teacher recorded the words and sentences spoken by the characters in the learning tasks, which generated a sense of reality similar to communication with a real person. Furthermore, when users pointed at a single word or finished arranging a sentence, the system provided corresponding pronunciation at the same time to strengthen the students' listening and communicative ability.
- Real-person motion capture: The system motion-captured every motion the user made in the learning task, such as waving their hands or nodding their heads, to simulate real situations.
- Room scale: With HTC VIVE's 360-degree tracking technology, users with headsets and controllers could physically walk around within a 7' x 7' play area. They could walk freely across the streets or enter into a room with their real-time motion reflected in the VR environment.
- Object taking: In the system, users could easily manipulate objects in the VR environment, such as picking up a toy car from a shelf or throwing trash into a trash can, with natural movement.
- Body tracking interaction: The system, integrated with spatial concepts and limb learning, provided an interactive interface that enabled users to use body language to communicate with others through the controller. For instance, if users did not know how to say the color "brown", they could simply point at or pick up a "brown object" instead.
- Flexible answering mode: The conversations in the system simulated real life situations; therefore, as long as the intended meaning was communicated, minor grammatical errors were acceptable. For example, when a user wanted to buy a toy car, a complete sentence such as "I would like to buy a toy car" or a brief phrase such as "buy a toy car" was considered acceptable in the learning system.
- Pointing to select: Due to the difficulty of identifying every user's voice, users in the system selected words to arrange a sentence to demonstrate their answers.

The sample task procedure of purchasing a toy is shown in Figure 1.

In this system, the students chose from several modes with variations in language level (easy or difficult), caption options (caption-on or caption-off), and an optional function for a time limit. After the initial setup, the students faced a task guided by an avatar. The task of purchasing a toy, for example, required the player to visit a place and respond to prompts initiated by the system avatar (see Figure 2 and 3). The mission clearly specified the kind, color, and brand of the toy the students should purchase (Figure 3). To accomplish the mission for communicative purposes, the students needed to understand the clerk's questions and use comprehensible sentences to express the need (see Figure 4 and 5). In addition to arranging different words into a sentence, the students recorded their responses into the system and picked up the objects in the system (see Figure 6 and 7).



***Task**: The learner needs to buy a toy for his/her little brother with random requirements (e.g., brand, type, color, price…)*

Figure 1. Task procedure in the system

Figure 2. Environment in VRLE

Figure 3. Mission in VRLE

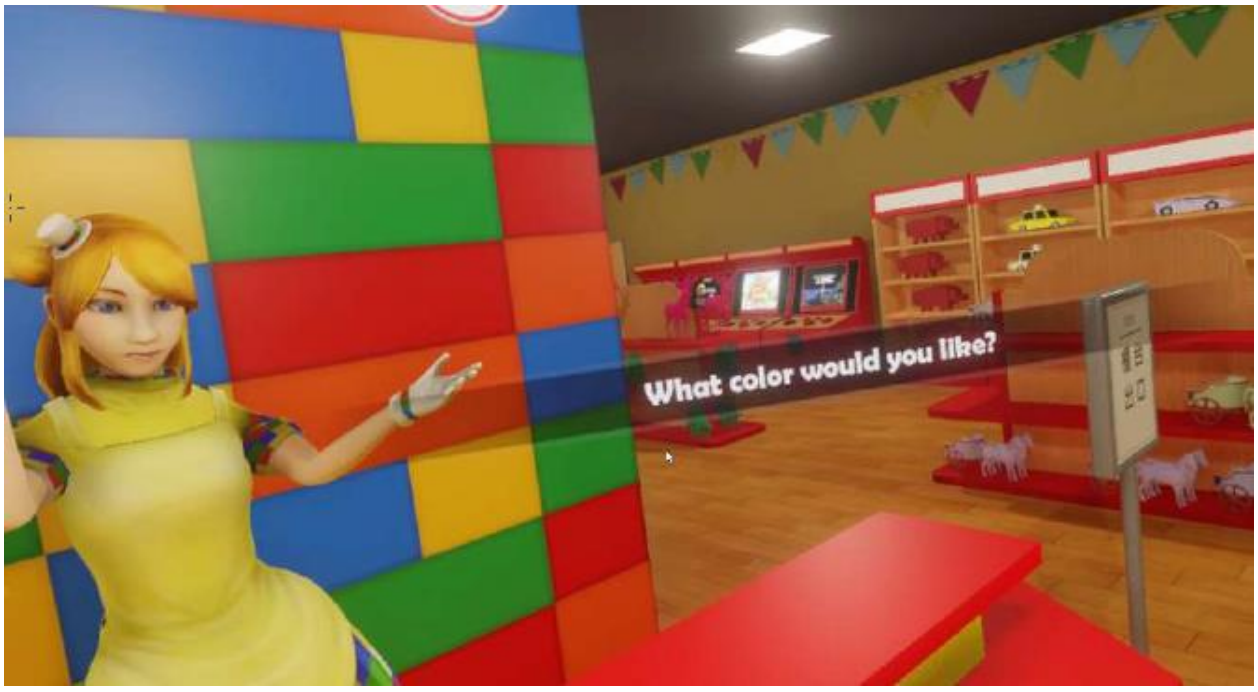


Figure 4. Communicative content in VRLE

Figure 5. Arranging different words into a sentence

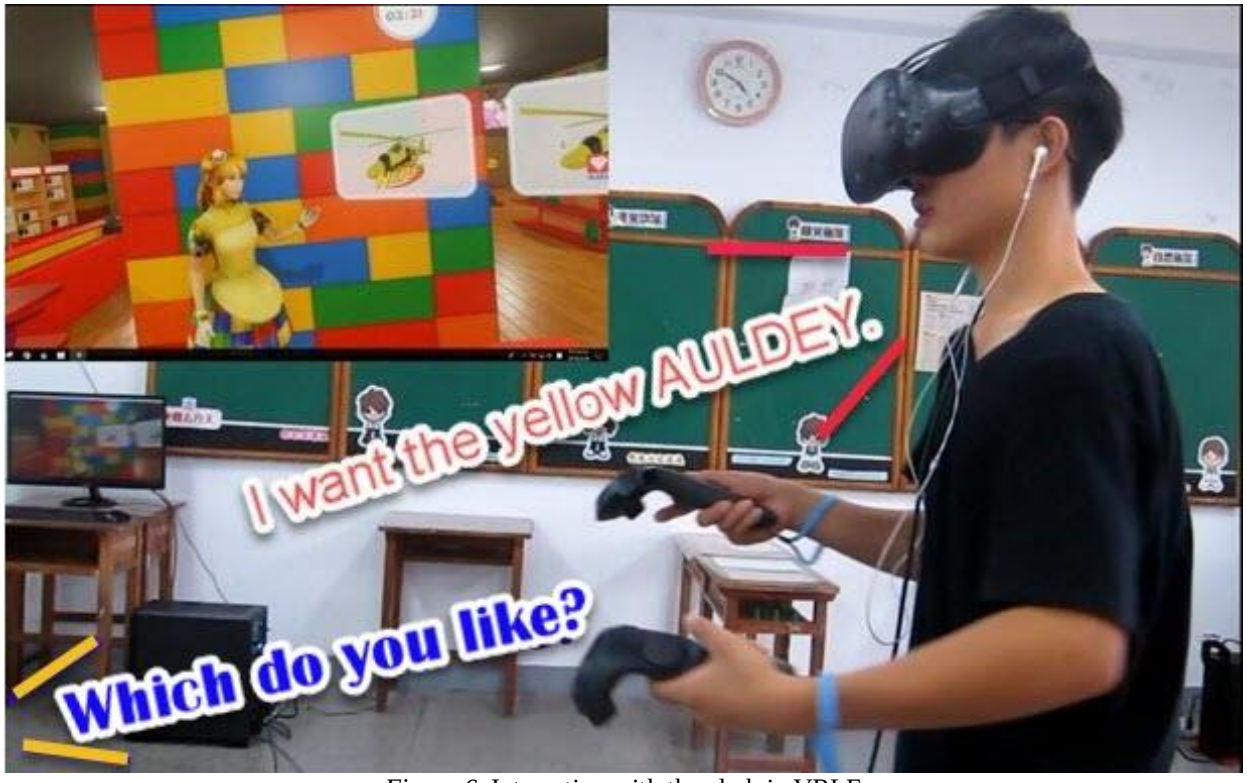


Figure 6. Interacting with the clerk in VRLE

Figure 7. Picking up the object

3.3. Research design

An affordance-based research design was adopted in this study, rather than an experimental/control group methodology. Affordances are "the qualities or properties of an object that define its possible uses or make clear how it can or should be used" (Merriam-Webster, N.D.). In teaching English, affordances define the capabilities and benefits that a teaching method or tool offers to instructional designers and students. When evaluating individual choices about learning technology, the affordances can be thought of as the effectiveness with which students can perform individual learning tasks (Marek & Wu, 2019).

The rationale for the affordance-based design of the current study originated from Colpaert's (2012) recommendation in his invited lecture for doctoral students that it is more valuable to study the affordances of a particular learning tool than to simply consider the differences between using or not-using the tool. He observed that so many different factors affect language learning that it is hard to predict whether the successful implementation of a technology-enhanced instructional design at one school will yield the same positive result at another school. In addition, because research participants are often students, there is a growing ethical concern about withholding learning experiences from some students in order to preserve a control group (Deygers, 2019).

Therefore, all of the participants in the current study experienced the VRLE system. The researchers geared data collection to understand the affordances of the system for teaching and learning, as embodied by the research questions about the communicative ability, sense of presence, and perceptions of affordances acquired by low-achieving EFL learners.

The VRLE included two modes, with one caption-assisted and the other without captions. The students first experienced the caption-on mode (Figure 8) and then the caption-off mode (Figure 9). It should be noted that while the system allowed the students to choose caption options, this study did not focus on how the sequence of caption provision affected the students' learning outcomes and perceptions. Rather, the purpose was to explore the students' perspectives about the caption mechanism in the VRLE system. To these low-achieving students, focusing on the learning tasks while at the time adapting themselves to the new learning system might make the students cognitively overloaded and lead to potential resistance to use the system. Therefore, to avoid confounding effects from different caption sequence designs on learning outcomes and perceptions, all of the students experienced the caption-on mode first and then the caption-off mode.

Figure 8. Caption-on mode in VRLE

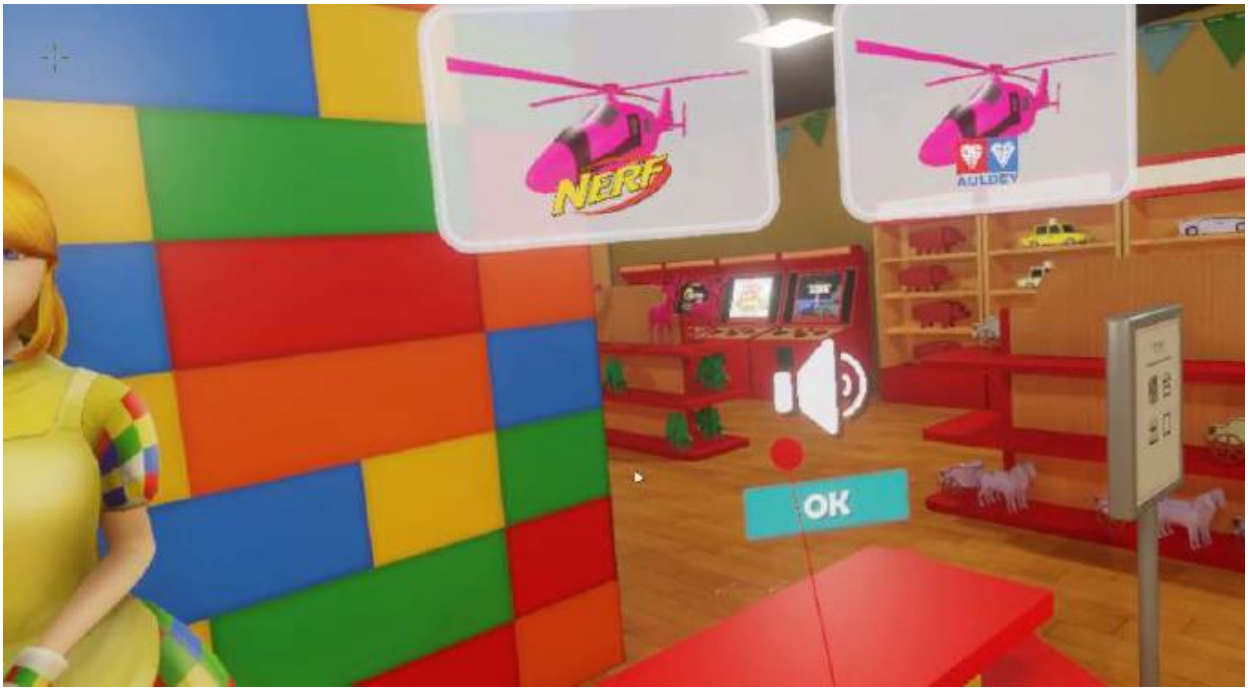


Figure 9. Caption-off mode in VRLE

3.4. Data collection and analysis

In response to RQ1 addressing the extent to which the VRLE system facilitated the communicative ability of EFL low-achieving learners, all of the students took an English pre-test focusing on communicative ability. It is worth noticing that to these low-achieving learners suffering from English instruction disadvantages, improvement in English learning meant basic abilities to get meaning across and to respond to linguistic cues in a communicative context. The researcher-developed pre-test assessed student communicative ability via listening and dialogic interaction, with each part containing 10 questions. For the listening part, the students listened to two people

conversing with each other (e.g., greetings, telephone chat, or weather of the day) and then rearrange those lines (around four to six) in correct order. As regards the dialogic interaction, the students listened to question prompts (e.g., "What is your favorite color?", "How is your day today?", "How's the weather today? Do you like it?") and made responses. The instructor and the researchers (also experienced teachers offering English instruction at universities) then graded the students' performances.

During data collection, the students were grouped into pairs by the teacher and each pair used the VR learning system outside the regular class time. All of the students experienced the four learning tasks embedded in VRLE, including making reservations at hotels and restaurants, purchasing a toy, asking for directions, and ordering a meal. Considering the potential side effects of dizziness and disorientation, each task lasted for around 15 minutes. Therefore, each student experienced the VRLE for around one hour in total within the four-week VR experience. Each students' VR learning experience was recorded on video for later examination. Right after the experiment, all of the students completed an English post-test. The post-test was identical to the pre-test in form, including listening and dialogic interaction. A paired-sample *t*-test examined whether significant differences existed between the pre-test and the post-test.

To address RQ2, the students responded to the Igroup Presence Questionnaire (IPQ) adapted from Schubert, Friedmann and Regenbrecht (2001) after the post-test, concerning the level of presence perceived in the VR learning setting. The original IPQ was comprised of 14 items rated in the form of a 7-point Likert Scale. The IPQ was used as a composite measure of presence with scores ranging from 14 to 98 and was divided into constructs assessing the four components of presence: overall feeling of presence, spatial presence, involvement, and realness. The questionnaire started with one item assessing the overall feeling of presence that the students perceived while using the VRLE system. Then, the construct of spatial presence contained five items assessing feelings that one was physically present within a virtual environment. The construct of involvement contained four items assessing attention to the virtual world. The construct of realness included four items assessing how real the virtual stimuli appeared. In addition, to explore whether different multimedia designs (i.e., caption-on and caption-off mode) affected the sense of presence, the researchers added the construct of caption-related presence with four questions, thus expanding the IPQ into an 18-item questionnaire.

Finally, to answer RQ3, about how the students perceived the use of VR for English learning, the students were invited to a semi-structured face-to-face interview. In answering, the participants (1) reflected on if the system provided an authentic setting, (2) compared the caption-on mode with the caption-off mode, and (3) made suggestions concerning how the system could be improved. The qualitative data was read repeatedly by the researchers and grouped into themes that recurred frequently. The researchers also analyzed their own notes on the experiences of the students for insights.

4. Results

The findings are organized in accordance with the research questions.

4.1. RQ1: To what extent did VRLE facilitate the communicative ability of EFL low-achieving learners?

To answer whether the use of the VRLE system in the experiment facilitated the low-achieving students' English communicative ability, descriptive analysis and a paired-samples *t*-test comparing the pre-test and the post-test were employed. Inter-rater reliability was measured with Krippendorff's alpha at .86, which is above the level considered the norm for good reliability (Hayes & Krippendorff, 2007). The results in Table 2 revealed that the mean score of the post-test (M = 68.82) was higher than that of the pre-test (M = 60.56). Further analysis using the paired-samples *t*-test (shown in Table 3) suggested that the students' performance on the post-test was significantly higher than that on the pre-test ($p < .001$), thus suggesting the facilitative role of the VRLE system on the communicative ability of the low-achieving participants.

Table 2. Descriptive statistics of the pre-test and post-test

Test	*N*	Mean	*SD*
Pre-test	72	60.56	23.16
Post-test	72	68.82	22.77

Table 3. Paired-samples *t*-test of the pre-test and post-test

	Paired differences					*t*	*df*	*Sig.* (2-tailed)
	Mean	*SD*	Std. error mean	95% *CI* Lower	Upper			
Pre-test – Post-test	-8.26	13.97	1.65	-11.55	-4.98	-5.02***	71	.000

Note. ***$p < .001$.

4.2. RQ2: To what extent did VRLE affect learners' presence in the virtual environment?

The responses of the students to the IPQ in Table 4 showed that the integration of VRLE for English communicative learning yielded an intermediate to upper-intermediate level of perception of presence, suggesting that most of the participants considered themselves as gaining a sense of presence in the VRLE learning environment. Among the five constructs, spatial presence (M = 5.31) topped the ranking, followed by overall feeling of presence (M = 5.27), caption-related presence (M = 4.48), involvement (M = 4.43), and realness (M = 3.88).

In terms of overall feeling of presence, the students generally expressed the feeling of being "in" the virtual environment when using the VRLE system, as evidenced by their responses to Item 1. The upper-intermediate level of perception regarding spatial presence (Items 2-6) suggested the students felt physically situated in the virtual space and that they felt a sense of action, revealing simulated authenticity of real-world scenarios in the VRLE system. As regards to the level of involvement (Items 7-10) in the VRLE system, while most of the students felt engaged in the virtual environment, they still experienced some interference from the real world, such as ambient sounds from their surroundings. In the construct of realness category (Items 11-14), the students' responses were slightly above average, indicating that the VRLE system might not be exactly like the real world despite the fact that the VR environment seemed consistent with the real-world experience. Finally, for caption-related presence (Items 15-18), the results were mixed. While most students expressed higher involvement in the caption-on mode than the caption-off mode, the caption-off mode made the VRLE system more authentic in its representation of the real world. The overall results of the study revealed that the VRLE system created an immersive environment, since the students felt situated in the virtual setting. In addition, the students gained a sense of involvement in the VR-based learning tasks while interacting with the virtual character. Last but not the least, while the caption assistance led the students to be more deeply immersed in the virtual environment, the system without captions was perceived to be more realistic.

Table 4. Descriptive statistics of Igroup presence questionnaire

Subscale and questionnaire item	Mean
Overall feeling of presence (M = 5.27)	
1. In the computer-generated world I had a sense of "being there."	5.27
Spatial presence (M = 5.31)	
2. Somehow I felt that the virtual world surrounded me.	5.30
3. I felt like I was just perceiving pictures.	5.57
4. I did not feel present in the virtual space.	5.23
5. I had a sense of acting in the virtual space, rather than operating something from outside.	5.06
6. I felt present in the virtual space.	5.39
Involvement (M = 4.43)	
7. How aware were you of the real world surrounding while navigating in the virtual world? (i.e., sounds, room temperature, other people, etc.)	4.09
8. I was not aware of my real environment.	4.11
9. I still paid attention to the real world.	4.67
10. I was completely captivated by the virtual world.	4.84
Realness (M = 3.88)	
11. How real did the learning task seem to you?	3.59
12. How much did your experience in the virtual environment seem consistent with your real-world	4.44

experience?		
13.	How real did the virtual world seem to you?	3.7
14.	The virtual world seemed more realistic than the real world.	3.8
Caption-related presence (*M* = 4.48)		
15.	I am completely involved in the virtual world with captions.	4.94
16.	I am completely involved in the virtual world without captions.	3.99
17.	I feel like being in the real world when I situate the environment with captions.	4.26
18.	I feel like being in the real world when I situate the environment without captions.	4.72

4.3. RQ3: What were the learners' overall perceptions of the VRLE system?

The students' overall perceptions about the VR learning experience, collected via semi-structured interviews, were analyzed for identifying themes. Eight students volunteered for the interviews. The overall results revealed that the participants showed positive perceptions about the learning experience adopted in this study. Their responses highlighted their perceptions about the VRLE system in terms of realness, engagement, perspectives on caption assistance, and system recommendations. However, some students also directed the attention to concerns that should be taken into account while using the VRLE system for learning.

4.3.1. Realness in the VR learning experience

Most of the students expressed that they seemed to be truly inside the virtual world. Being able to move and take VR objects freely made the whole experience more realistic, enhancing their engagement in the VR system. In addition, the fact that the voice of the virtual character was recorded by a native English speaker created a sense of actually communicating with a real person. Related responses from the students were as follows:

- "I think it is quite real! When I moved too close to the cabinet, I felt like I was going to bump my head on it." (A1)
- "The things in VRLE are really realistic, but the movements of the clerk make it feel unreal because people in the real world won't act like that." (A4)
- "I think the clerk is real because her accent is not like a Taiwanese. I felt I was really talking to a foreigner." (A5)
- "I think the toys in VRLE are super real and I think it is cool that I could pick up items freely or throw away things in the virtual world." (A6)

However, some students also expressed their concerns while experiencing the VRLE system. Some students complained about experiencing discomfort in the VRLE system. One student even mentioned the feeling of "getting lost and disoriented," while another stated dizziness that prevented him from appropriately completing the task. A few students complained about challenges faced in grabbing the objects in the system, despite the time given for them to practice before actual experiment implementation.

4.3.2. Engagement in the VR learning

As the students felt a sense of realness from the VRLE system, their engagement in the VR system grew. Once the student felt engaged, learning took place naturally, as evidenced by the students' attentiveness to the learning tasks embedded in the VRLE system.

- "I felt like I really stayed inside and I concentrated on answering the questions." (A1)
- "I completely focused on the learning task! It is so rare that I listen to English so carefully!" (A7)

Similar to the issue of realness in the VR, some students noted their cautions with the use of VRLE system. One student mentioned the disturbance by the wire connected to the headset, making him "afraid of being tripped over." Some students, on the other hand, pointed out the anxiety of losing face in front of their classmates.

4.3.3. Perspectives on the caption

Some students preferred the caption-on mode, since they could cross verify their comprehension. To students who were still learning how to use English for daily communicative purposes (low-achieving students, in this study), the use of captions made them feel secured. On the other hand, some students were fonder of the caption-off mode, since it removed the visual distraction and thus created a more realistic setting.

- "I prefer to have the captions on since I can have a better understanding of the conversation. It was harder for me to understand what the clerk said without captions." (A1)
- "I like the caption option because I knew what the clerk was talking about; however, I think I can learn more without captions. It would push me to stay focused on the listening." (A7)
- "I like the captions. If I can't read the captions, I feel a bit lost. I need to double check my comprehension with the captions." (A4)
- "I prefer no captioning because it is more realistic. I can also listen to what the clerk said more attentively and that helps me learn better." (A2)
- "I like the no caption option, because the caption covered my sight and made me dizzy. No caption is more effective to me." (A3)
- "I think it's more effective to learn without captions because it is more like a real conversation! However, when it comes to breaking the record, the assistance of captions would be necessary." (A8)

Overall speaking, the students' preference for the caption assistance was mixed, with some favoring the caption-on mode while the others preferred the caption-off mode. For some students, the inclusion of the caption in the VRLE system made them feel secured as it provided extra linguistic support. To the others, the exclusion of the caption enhanced the sense of realness and authenticity in the VRLE system. Such results also echoed their responses of the caption-related presence subscale of the IPQ.

4.3.4. Recommendations for the system

While the students were positive about the use of VRLE in enhancing their communicative ability, suggestions were offered from the participants to improve the overall quality of the VR learning system. These recommendations included adding more stages or providing hints/rewards while using the system. Other comments to the systems were as follows:

- "I wish the playing area could've been bigger, such as walking outside to enter other stores." (A1)
- "I think there should be other people in the store—for example, asking other people questions or competing with other players." (A2)
- "I think it would be more realistic if I was able to see my own feet, or part of my body." (A3)
- "I wish that there was a reward mechanism so that players can collect points by answering questions and receive virtual money as a reward." (A5)
- "I think there should be a hint mechanism which allows you to get hints among the objects in the area, such as posters or TV. Otherwise it was quite annoying to get stuck on the same question." (A6)

Taken together, the students' suggestions on the VRLE system focused on more contents to be included, interaction with more avatars, improvement on the first-person physical appearance design, and inclusion of hints/rewards mechanism.

5. Discussion

The overall findings of the study showed that the VRLE system could provide the communication-facilitator affordances necessary to address the requirements in the academic literature for beneficial VR-supported language learning and of the more limited literature on VR for communicative purposes, particularly to EFL low-achieving learners. Specifically, the VR-supported design, using VRLE, contributed to significantly higher learning outcomes among those low-achieving students and most of them were positive about the VRLE system regarding spatial presence, involvement, and a sense of realness. They also found VRLE to be a beneficial tool that facilitated the communicative aspect of language learning, but their perceptions about the caption provision were more mixed.

5.1. Contextualized learning in VR to facilitate language learning

The results revealed significant differences in the growth of learners' communicative ability before and after using the VRLE system, potentially suggesting the facilitative role that the VRLE system played on the communicative ability of the low-achieving participants. The results were in line with the study of Legault et al. (2019) that less successful learners exhibited greater gains via immersive VR in second language learning. Further, students were not left alone to make the connection of language skills and applications, since the system itself provided an authentic immersive environment. It is not surprising that the students made improvement as the learning tasks embedded in the VRLE system reflected John Dewey's philosophy of learning by doing, indicating that the hands-on mission the students experienced in the virtual environment enabled them to interact with the system and to adapt as well as learn. As a high-immersion VR system that involved egocentric navigation rather than exocentric navigation commonly seen in low-immersion VR systems (Kozhevnikov & Dhond, 2012), the VRLE system offered a simulated real-life scenario for learners to test and apply language skills as a whole. It echoed the statement by Shu, Huang, Chang, and Chen (2019) that head-mounted display (such as the VRLE system adopted in this study) offered a greater sense of presence in the contextualized learning setting, indicating a positive potential for language learning. Unlike conventional spoken or written assessments where language skills of grammar, articulation, speaking, listening, reading and writing are tested as a detached skill, the learning system adopted in the study provided task-based assignments for the students and integrated communicative ability, thus making language learning holistic, rather than isolated aspects to be mastered separately (Robinson, 2011). Furthermore, by observing the performance of their classmates, the students in this study were given opportunities to observe and imitate behaviors of their peers, with which they tried to improve their learning outcomes.

5.2. Immersive learning through VR to attain engagement and lower level of anxiety

The goal of creating a sense of presence in the virtual world was achieved in this study, leading to an effective immersion experience for language acquisition among the students, aligning with the positive beliefs about immersion in language learning (see Paige, Jorstad, Siaya, Klein, & Colby, 2000 for a review of literature). In fact, this pedagogical approach aims for learners to maintain constant contact with the target language, which would be particularly beneficial to EFL learners who do not have easy access to the linguistic and cultural elements of the target language (Freina & Ott, 2015). Learning a second language (L2) through a real-life immersive environment, namely learning in the target culture where the language is spoken in real life, leads to lower interferences from a learner's native language to L2 and yields to higher proficiency compared with learning in a conventional classroom setting (Legault et al., 2019; Linck et al., 2009). Nevertheless, real-life immersion for various scenarios is not always accessible to every L2 learner, not to mention to the students in this experiment who are learning in a disadvantaged English learning environment. With the help of immersive VR technology, rendering capable the recreation of immersive learning settings, language learners are better surrounded with simulated environments that might not have been so easily accessible in the past. Furthermore, VR with HMD also enables a higher degree of embodiment in a virtual setting which proves to be conducive to L2 learning (Legault et al., 2019). The VRLE system adopted in this research offered learners a chance to simulate interactions in a real-life scenario. In addition, immersive learning offers language learners self-directed exploration rather than conventional spoon-fed and lecture-based instruction, which often disengages students (Delialioğlu, 2012). Accordingly, learners are empowered with the capability to self-direct learning, thus contributing to increased ownership of learning and engagement (Rashid & Asghar, 2016). Furthermore, the VR-supported learning in this study led to a lowered level of anxiety among its users, echoing the potential benefit of integrating VR into language learning as indicated by Cheng and Tsai (2019) and Marquess et al. (2017). To be more specific, learners tend to have lowered affective filters while interacting with virtual characters, because they know they are interacting with a machine where taking risks in language production is encouraged (Lee & Chen Hsieh, 2019; Lee, Lee, & Chen Hsieh, 2019; Reinders & Wattana, 2015). That is, VR-supported technology saves learners the embarrassment of making mistakes, hence increasing overall student engagement specifically in language production skills.

5.3. Caption-on vs. caption-off in language learning

The students' perceptions about the caption design in the VRLE system suggested that the inclusion of captions enabled them to be more involved in the VR learning, despite the fact that the exclusion of captions actually made

the learning experience more realistic. Previous studies have shown that supplying full captions or captions of target vocabulary in audio-visual materials has been an effective way to boost listening and reading comprehension of a second language (Hsu, Hwang, Chang, & Chang, 2013). The phenomenon of mixed feelings about captions might be explained by the sequence of the research design, with the caption-on mode played for the first round and then the caption-off mode for the second round. Learners were thus more acquainted with the interface, flow, tasks, etc. before experiencing the caption-off mode. Another potential reason, observed from the student interviews, is that captioning might sometimes distract the attention of users from their assigned tasks. While some students rely more on captions to cross-validate their comprehension, other students were immersed in the learning task and thus might perceive captions as a distraction, which potentially hindered the realness of the VR design.

6. Conclusion

The results of this study have extended prior research by probing into the under-explored issue of using virtual reality for communicative purposes in EFL learning, particularly regarding low-achieving junior high school students in a disadvantaged English learning environment. The primary affordances identified by this study were the beneficial applications of the VRLE system toward English communicative ability and sense of presence in an EFL context. Based on the findings and discussion of this study, the researchers offer the following conclusions and recommendations for practice.

- VR-supported instruction is an appropriate pedagogical design for teaching communicative aspects of English, since it aligns with modern ideas of student-centered active learning (Nouri, 2016), enables low-achieving learners to be immersed and motivated in learning tasks, and leads to beneficial outcomes.
- Because the effects of multimedia design on caption-on and caption-off modes were more mixed, instructors should take into consideration curricular goals and student needs. While the caption-on mode could increase student understanding of the materials, thus enhancing the level of involvement, the caption-on mode, on the other hand, might lower the realness of the virtual learning. Instructors are therefore advised to tailor the adoption of captioning to address different individual learning needs.

The present study not only provided empirical evidence for a VR-supported learning context among EFL low-achieving learners, but also shed light on the scenario of technology-enhanced, innovative pedagogies. Based on the results of this study, it is essential to conduct follow-up studies that would address the challenges raised in this study—that is, easy and affordable access to VR learning, and VR systems for more aspects of English learning. While VR-supported learning has been shown to be effective in this study, the high cost of the equipment makes it challenging to be widely adopted in classroom settings on a regular basis (McFaul & FitzGerald, 2020). Future design on VR learning systems, therefore, could combine VR with applications on smartphones to make VR learning more accessible. Furthermore, the VR system developed in the study focused on enhancing English communicative ability—namely, how to encourage students to apply their communicative skills purposefully in real-life scenarios where minor grammar mistakes may not affect comprehension. Therefore, it may not be the most appropriate model for other language skills such as grammar or writing practices. More VR systems are needed so that a more comprehensive understanding of how VR facilitates EFL learning can be achieved.

7. Final thoughts

The use of high-immersion VR in EFL contexts has been promising because VR is able to provide a near-authentic contextualized environment while also allowing for meaningful language engagement and promoting learner-centered approaches. In turn, this activates intrinsic motivation to lifelong language learning. The educational application of VR offers a whole new arena for language learners as it enables learners to be perceptually inside a scenario and to apply English skills virtually beyond the traditional classroom walls.

References

Abed, A. Q. (2011). Teachers' awareness of second language learning strategies adopted and used by their students: A Questionnaire. *Theory and Practice in Language Studies, 1*(3), 205-218.

Barrot, J. S. (2016). Using Facebook-based e-portfolio in ESL writing classrooms: Impact and challenges. *Language, Culture and Curriculum, 29*(3), 286-301.

Canale, M. (2014). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and Communication* (pp. 14-40). New York, NY: Routledge.

Chang, S. C., Hsu, T. C., & Jong, M. S. Y. (2020). Integration of the peer assessment approach with a virtual reality design system for learning earth science. *Computers & Education, 146*, Article 103758. doi:10.1016/j.compedu.2019.103758

Checa, D., & Bustillo, A. (2020). A Review of immersive virtual reality serious games to enhance learning and training. *Multimedia Tools and Applications, 79*, 5501-5527.

Chen Hsieh, J. S., Wu, W. C. V., & Marek, M. W. (2017). Using the flipped classroom to enhance EFL learning. *Computer Assisted Language Learning*, *30*(1-2), 1-21.

Chen, J. C. (2016). The Crossroads of English language learners, task-based instruction, and 3D multi-user virtual learning in Second Life. *Computers & Education, 102*, 152-171.

Cheng, K.-H., & Tsai, C.-C. (2019). A Case study of immersive virtual field trips in an elementary classroom: Students' learning experience and teacher-student interaction behaviors. *Computers & Education. 140*, Article 103600. doi:10.1016/j.compedu.2019.103600

Chien, S. Y., Hwang, G. J., & Jong, M. S. Y. (2020). Effects of peer assessment within the context of spherical video-based virtual reality on EFL students' English-Speaking performance and learning perceptions. *Computers & Education, 146*, Article 103751. doi:10.1016/j.compedu.2019.103751

Chou, M. H. (2018). Speaking anxiety and strategy use for learning English as a Foreign Language in full and partial English-medium instruction contexts. *TESOL Quarterly*, *52*(3), 611-633.

Colpaert, J. (2012, May). *The Role of personal goals in designing learning environments*. *Presented at the XVth International CALL Research Conference*. Providence University, Taichung, Taiwan.

Dawley, L., & Dede, C. (2014). Situated learning in virtual worlds and immersive simulations. In *Handbook of research on educational communications and technology* (pp. 723-734). New York, NY: Springer.

Deygers, B. (2019). Fairness and social justice in English language assessment. In X. Gao (Ed.), *Second Handbook of English Language Teaching* (pp. 541-569). Basil, Switzerland: Springer Nature Switzerland AG.

Delialioğlu, Ö. (2012). Student engagement in blended learning environments with lecture-based and problem-based instructional approaches. *Educational Technology and Society, 15*, 310-322.

Diemer, J., Alpers, G. W., Peperkorn, H. M., Shiban, Y., & Mühlberger, A. (2015). The Impact of perception and presence on emotional reactions: A review of research in virtual reality. *Frontiers in Psychology, 6*, 26.

Dolgunsöz, E., Yildirim, G., & Yildirim, S. (2018). The Effect of virtual reality on EFL writing performance. *Journal of Language and Linguistic Studies, 14*(1), 278-292.

Dörnyei, Z. (2014). Researching complex dynamic systems: "Retrodictive qualitative modeling" in the language classroom. *Language Teaching, 47*, 80-91.

Estes, J. S., Dailey-Hebert, A., & Choi, D. H. (2016). Integrating technological innovations to enhance the teaching-learning process. In D. Choi, A. Dailey-Hebert, & J. S. Estes (Eds.), *Emerging Tools and Applications of Virtual Reality in Education* (pp. 277-304). Hershey, PA: IGI Global.

Freina, L., & Ott, M. (2015). A Literature review on immersive virtual reality in education: State of the art and perspectives. In *Proceedings of the 11th International Scientific Conference on eLearning and Software for Education* (pp. 133-141). Bucharest, Romania: Advanced Distributed Learning Association.

González-Lloret, M., & Ortega, L. (2014). *Technology-mediated TBLT: Researching technology and tasks*. Amsterdam, The Netherlands: John Benjamins.

Hassani, K., Nahvi, A., & Ahmadi, A. (2016). Design and implementation of an intelligent virtual environment for improving speaking and listening skills. *Interactive Learning Environments, 24*(1), 252-271.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*, 77-89.

Hsu, C. K., Hwang, G. J., Chang, Y. T., & Chang, C. K. (2013). Effects of video caption modes on English listening comprehension and vocabulary acquisition using handheld devices. *Educational Technology & Society, 16*(1), 403-414.

Huang, H. M., Rauch, U., & Liaw, S. S. (2010). Investigating learners' attitudes toward virtual reality learning environments: Based on a constructivist approach. *Computers & Education*, *55*(3), 1171-1182.

Hu-Au, E., & Lee, J. J. (2017). Virtual reality in education: A Tool for learning in the experience age. *International Journal of Innovation in Education*, *4*(4), 215-226.

Hymes, D. (1972). On communicative competence. In J. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Harmondsworth, England: Penguin Books.

Kaplan-Rakowski, R., & Gruber, A. (2019). Low-immersion versus high-immersion virtual reality: Definitions, classification, and Examples with a foreign language focus. In *Proceedings of the 12th International Conference Innovation in Language Learning* (pp. 552-555). Florence, Italy: Filodiritto Editore.

Kozhevnikov, M., & Dhond, R. P. (2012). Understanding immersivity: Image generation and transformation processes in 3D immersive environments. *Frontiers in Psychology*, *3*, 284.

Lan, Y. J. (2014). Does second life improve mandarin learning by overseas Chinese students? *Language Learning & Technology*, *18*(2), 36-56.

Lan, Y. J. (2015). Contextual EFL learning in a 3D virtual environment. *Language Learning & Technology*, *19*(2), 16-31.

Lanier, J. (2017). *Dawn of the new everything: A Journey through virtual reality*. London, UK: Random House.

Lee, E. A.-L., & Wong, K. W. (2014). Learning with desktop virtual reality: Low spatial ability learners are more positively affected. *Computers & Education*, *79*, 49-58.

Lee, J. S., & Chen Hsieh, J. (2018). University students' perceptions of English as an International Language (EIL) in Taiwan and South Korea. *Journal of Multilingual and Multicultural Development*, *39*(9), 789-802.

Lee, J. S., & Chen Hsieh, J. (2019). Affective variables and willingness to communicate of EFL learners in in-class, out-of-class, and digital contexts. *System*, *82*, 63-73.

Lee, J. S., Lee, K., & Chen Hsieh, J. (2019). Understanding willingness to communicate in L2 between Korean and Taiwanese students. *Language Teaching Research*. doi:10.1177/1362168819890825

Lee, K. M. (2004). Presence, explicated. *Communication Theory*, *14*(1), 27-50.

Legault, J., Zhao, J., Chi, Y. A., Chen, W., Klippel, A., & Li, P. (2019). Immersive virtual reality as an effective tool for second language vocabulary learning. *Languages*, *4*(1), 13. doi:10.3390/languages4010013

Levak, N., & Son, J. B. (2017). Facilitating second language learners' listening comprehension with Second Life and Skype. *ReCALL*, *29*(2), 200-218.

Lin, T. J., & Lan, Y. J. (2015). Language learning in virtual reality environments: Past, present, and future. *Educational Technology & Society*, *18*(4), 486-497.

Linck, J. A., Kroll, J. F., & Sunderman, G. (2009). Losing access to the native language while immersed in a second language: Evidence for the role of inhibition in second-language learning. *Psychological Science*, *20*(12), 1507-1515.

Luo, W. H. (2017). Teacher perceptions of teaching and learning English as a lingua franca in the expanding circle: A Study of Taiwan: What are the challenges that teachers might face when integrating ELF instruction into English classes? *English Today*, *33*(1), 2-11.

Makransky, G., & Lilleholt, L. (2018). A Structural equation modeling investigation of the emotional value of immersive virtual reality in education. *Educational Technology Research and Development*, *66*(5), 1141-1164.

Marek, M. W., & Wu, W-C. V. (2014). Environmental factors affecting computer assisted language learning success: A Complex dynamic systems conceptual model. *Computer Assisted Language Learning*, *27*(6), 560-578.

Marek, M. W. & Wu, W-C. V. (2019). Creating a technology-rich English language learning environment. In X. Gao (Ed.), *Second Handbook of English Language Teaching* (pp. 757-777). Springer Nature Switzerland AG: Basil, Switzerland.

Marquess, M., Johnston, S. P., Williams, N. L., Giordano, C., Leiby, B. E., Hurwitz, M. D., Dicker, A. P., & Den, R. B. (2017). A Pilot study to determine if the use of a virtual reality education module reduces anxiety and increases comprehension in patients receiving radiation therapy. *Journal of Radiation Oncology*, *6*(3), 317-322.

Matsuda, A. (Ed.). (2017). *Preparing teachers to teach English as an international language*. Bristol, UK: Multilingual Matters.

McFaul, H., & FitzGerald, E. (2020). A Realist evaluation of student use of a virtual reality smartphone application in undergraduate legal education. *British Journal of Educational Technology*, *51*(2), 572-589.

McKay, S. L., & Brown, J. D. (2016). *Teaching and assessing EIL in local contexts around the world*. New York, NY: Routledge.

Merriam-Webster. (N.D.). *Affordance: Definition of affordance by Merriam-Webster*. Retrieved from https://www.merriam-webster.com/dictionary/affordance

Nouri, J. (2016). The Flipped classroom: For active, effective and increased learning–especially for low achievers. *International Journal of Educational Technology in Higher Education, 13*(1), 33. doi:10.1186/s41239-016-0032-z

Paige, R. M., Jorstad, H., Siaya, L., Klein, F., & Colby, J. (2000). Culture learning in language education: A Review of the literature. In R. M. Paige, D. L. Lange, & Y. A. Yeshova (Eds.), *Culture as the Core: Integrating Culture into Second Language Curriculum* (pp. 173-236). Minneapolis, MN: University of Minnesota.

Pham, V. P. H., & Usaha, S. (2016). Blog-based peer response for L2 writing revision. *Computer Assisted Language Learning, 29*(4), 724-748.

Piaget, J. (1969). The Intellectual development of the adolescent. In G. Caplan & S. Lebovici. (Eds.), *Adolescence: Psychosocial Perspectives* (pp. 22-26). New York, NY: Basic Books.

Rashid, T., & Asghar, H. M., (2016). Technology use, self-directed learning, student engagement and academic performance: Examining the interrelations. *Computers in Human Behavior, 63*, 604-612.

Reinders, H., & Wattana, S. (2015). Affect and willingness to communicate in digital game-based learning. *ReCALL, 27*(1), 38-57.

Rivers, W. M. (1972). Talking off the tops of their heads. *TESOL Quarterly*, *6*(1), 71-81.

Robinson, P. (2011). *Task-based language learning*. Ann Arbor, MI: Language Learning Research Club, University of Michigan.

Rouhi, A., & Saeed-Akhtar, A. (2008). Planning time: A mediating technique between fluency and accuracy in task-based teaching. *Journal of English Language Pedagogy and Practice, 1*(Inaugural Issue), 103-133.

Sadler, R., Dooly, M., Thomas, M., Reinders, H., & Warschauer, M. (2013). Language learning in virtual worlds: Research and practice. In M. Thomas, H. Reinders, & M. Warschauer (Eds.), *Contemporary Computer Assisted Language Learning* (pp. 159-182). New York, NY: Bloomsbury.

Schubert, T., Friedmann, F., & Regenbrecht, H. (2001). The Experience of presence: Factor analytic insights. *Presence-Teleoperators and Virtual Environments, 10*(3), 266-281.

Shu, Y., Huang, Y. Z., Chang, S. H., & Chen, M. Y. (2019). Do virtual reality head-mounted displays make a difference? A Comparison of presence and self-efficacy between head-mounted displays and desktop computer-facilitated virtual environments. *Virtual Reality*, *23*(4), 437-446.

Usoh, M., Catena, E., Arman, S., & Slater, M. (2000). Using presence questionnaires in reality. *Presence: Teleoperators & Virtual Environments, 9*(5), 497-503.

Vrellis, I., Avouris, N., & Mikropoulos, T. A. (2016). Learning outcome, presence and satisfaction from a science activity in Second Life. *Australasian Journal of Educational Technology, 32*(1), 59-77.

Vygotsky, L. (1978). Interaction between learning and development. *Readings on the Development of Children, 23*(3), 34-41.

Wang, C. X., Anstadt, S., Goldman, J., & Mary, L. M. (2014). Facilitating group discussions in Second Life. *Journal of Online Learning and Teaching, 10*(1), 139-152.

Wang, W. C., Lin, C. H., & Lee, C. C. (2011). Thinking of the textbook in the ESL/EFL classroom. *English Language Teaching, 4*(2), 91-96.

Wang, Y. F., Petrina, S., & Feng, F. (2017). VILLAGE—Virtual immersive language learning and gaming environment: immersion and presence. *British Journal of Educational Technology*, *48*(2), 431-450.

Witmer, B. G. & Singer, M. J. (1998). Measuring presence in virtual environments: A Presence questionnaire. *Presence: Teleoperators and Virtual Environments, 7*(3), 225-240.

Wu, W. C. V., Yen, L. L., & Marek, M. (2011). Using online EFL interaction to increase confidence, motivation, and ability. *Educational Technology & Society, 14*(3), 118-129.

Zhang, Y., & Liu, L. (2018). Using computer speech recognition technology to evaluate spoken English. *Kuram ve Uygulamada Egitim Bilimleri*, *18*(5), 1341-1350.

Zou, B., Wang, D., & Xing, M. (2016). Collaborative tasks in Wiki-based environment in EFL learning. *Computer Assisted Language Learning*, *29*(5), 1001-1018.

Student Game Design as a Literacy Practice: A 10-Year Review

Hsiu-Ting Hung[1], Jie Chi Yang[2*] and Yi-Chin Tsai[1]

[1]Department of English, National Kaohsiung University of Science and Technology, Taiwan // [2]Graduate Institute of Network Learning Technology, National Central University, Taiwan // hhung@nkust.edu.tw // yang@cl.ncu.edu.tw // 0431813@nkust.edu.tw

[*]Corresponding author

ABSTRACT: Learning through designing digital games has recently emerged as a potential approach for school learners to boost their literacy development and learning in and across disciplines. However, existing knowledge on this relatively new approach is still fragmented, and little is known about its implementation features, associated learning opportunities, and possible challenges experienced by students. As such, the present review seeks to synthesize relevant research in terms of the three aspects stated above to better understand the concept of student game design as a literacy practice. A total of 30 peer-reviewed research articles published between 2010-2020 are included in this research synthesis. Findings reveal that there is considerable variation in how the literacy learning approach of student game design is currently implemented, with respect to the school learners involved and game-making tools adopted. Despite its diverse nature, the feasibility of literacy learning by game-making is confirmed across the reviewed studies, with the disciplinary literacy in computer science and 21st century literacy being most prominent. This review has also brought to light the potential of introducing students to content-based game design to foster interdisciplinary learning. In order to provide a balanced portrait, this review further identifies major challenges of learning with the game-making approach from students' perspectives.

Keywords: Digital games, Game design, Literacy learning, Literature review

1. Introduction

There has been widespread recognition of the need for educators to re-conceptualize what it means to be literate, and how literacy learners can be educated to succeed in the 21st century (Mills, 2010; Trilling & Fadel, 2009). While the traditional notion of literacy centers on print-based practices of reading and writing, recent understanding of literacy is tightly linked to a repertoire of practices for functioning well in context-specific settings, which are mediated and shaped by technology in some way in this digital era (Gilster, 1997; Kress, 2003; Lankshear & Knobel, 2003). Arguably, literacy is now best understood as a broad range of socially organized practices that extend the traditional reading and writing skills. It follows that literacy can be practiced in varying forms for different purposes in a variety of sociocultural contexts, hence new literacies (Street, 1998) or multiliteracies (Cope & Kalantzis, 2000). Some notable examples addressed in this study are 21st century literacy (Trilling & Fadel, 2009), new media literacy with respect to digital game design (Buckingham & Burn, 2007), and disciplinary literacy in various subjects, such as computer science and social studies (Shanahan & Shanahan, 2008).

In line with the reconceptualization of literacy, educators and researchers are continually looking for innovative ways to help students learn effectively with digital technologies. Among the various options, digital games have been suggested as promising catalysts. In the book entitled "*What video games have to teach us about learning and literacy,"* Gee (2003) identifies 36 principles from cognitive science that are situated in games. For example, the active learning principle states that all aspects of the digital game-based learning environment are designed to encourage active student learning. Over the years, research interest in digital games has grown, and many of Gee's (2003) claims about the affordances of learning through game-playing have been supported by empirical studies. Research has shown that exposing students to well-designed gaming environments with appropriate instructional support can enhance their learning motivation (Hawlitschek & Joeckel, 2017), vocabulary acquisition (Franciosi, 2017), problem solving (Eseryel, Law, Ifenthaler, Ge, & Miller, 2014), and disciplinary literacy (Chen, Wong, & Wang, 2014). This strand of research promotes the approach to learning by playing digital games, which is taken as an initial effort to explore game-related applications in education (Boyle et al., 2016; De Freitas, 2018; Hung, Yang, Hwang, Chu, & Wang, 2018).

Building further on the educational potential of game use from a different perspective, a recent trend has seen the introduction of a game-making approach "in which games are designed by students (rather than professionals) for

learning benefits" (Kafai & Burke, 2015, p. 314). This approach is rooted in constructionist learning theories (Papert, 1980, 1991). It highlights the role of students as active learners as they take part in the process of constructing their own digital games (Prensky, 2008), and thereby constructing meaningful knowledge and experience for themselves (Kafai & Resnick, 2012). Various benefits of learning with the game-making approach have been shown in empirical studies, such as enhancing student game designers' creative thinking (Navarrete, 2013), improving their computer science knowledge and programming skills (Denner, Werner, & Ortiz, 2012), and actively engage them in the process of learning by design (Topalli & Cagiltay, 2018). Although there appears to be an increasing number of studies on student game design in recent years, this body of research is still small (Reynolds, 2016). Scholars have therefore called for more studies and reviews in order to more fully grasp the value of the game-making approach (Kafai & Burke, 2016; Kordaki & Gousiou, 2017). The present study is an endeavor in response to this call.

The purpose of this study is to provide a scoping review of empirical studies that adopt the game-making approach in educational contexts, using a content analysis of multiple aspects. Of central interest to this review are literacy practices of school learners across different levels of education, ranging from kindergarten to university (also known as K-16). Therefore, the first aspect analyzed here is student game designers' educational levels. This information is helpful to determine suitable settings for future implementations. Another related aspect is the existing tools for non-experts to design digital games for the sake of schooling. This is the practical information that allows educators to choose appropriate game design tools that best suit their target learners' needs. In addition to contextual features, researchers are generally interested in understanding what learning opportunities are available to students and what learning challenges are facing students as they are involved in the creation of digital games. While empirical evidence on the contributions and constraints related to the game-making approach is still inconclusive, review results of these aspects are intended to enhance the current knowledge base. Accordingly, the following research questions are analyzed in this review.

- What is known about the student game designers' educational levels and game design tools when learning with the game-making approach?
- What is known about the opportunities offered by the game-making approach for literacy learning?
- What is known about the challenges of learning with the game-making approach from students' point of view?

2. Related work

Several previous reviews informed this work. Li and Tsai (2013) reviewed 31 empirical studies published between 2000 and 2011 regarding the use of digital games in science education. The sample was identified through the databases of SCOPUS and Web of Science. The results revealed that most of the studies adopted the game-playing approach to facilitate students' science learning, and only two studies utilized the game-making approach for the same purpose. Learning gains in scientific knowledge were found to be the most dominant outcome, followed by problem solving skills.

Another related review by Kordaki and Gousiou (2016) was conducted in the context of computer science education. One of its main purposes was to examine the effects of a specific genre, digital card games, on student learning. Of the 24 articles spanning 2003-2013 that were located by database searches (e.g., ACM, ERIC, and IEEE), two-thirds asked students to learn by designing their own games, and one-third exposed students to game-playing environments for learning. Positive effects of both game uses were reported, with most centering on the acquisition of programming knowledge and skills.

In a follow-up study, Kordaki and Gousiou (2017) expanded the scope of their prior review from the domain-specific context of computer science to various application domains. A similar methodology was utilized to sample a total of 50 articles with varying game uses (game-making: $n = 14$; game-playing: $n = 35$; both: $n = 1$). The results provided evidence to support applications of digital card games in education in general, with computer science, language, and science being the most common disciplines.

Focusing on the learning benefits of the game-making approach, Kafai and Burke (2015) carried out a literature review to analyze research evidence on student game design in terms of personal, social, and cultural dimensions in K-12 education. The literature search sources included electronic databases, journal archives, and conference archives. Based on the review results of the 55 articles published in 1995-2015, it was found that the game-making

approach contributed most to one's growth in the personal dimension. Leading the way were studies that documented students' learning of coding, followed by the learning of other content areas, such as mathematics and science.

The four reviews mentioned above, although differing in focus, all agree in suggesting the feasibility of integrating digital games in education through the game-making approach. They have raised attention to the still evolving concept of student game design in various disciplinary contexts. To advance in this direction, the present review was motivated to synthesize research findings on the use of digital game design as a literacy practice for school learners.

Kafai and Burke's (2015) review is of particular relevance to this study. They synthesized research findings published up to 2015, and proposed a useful framework for organizing learning benefits of student game design along three different dimensions: personal, social, and cultural. Reflecting the continuing interest in the game-making approach, the present review attempts to provide a more up-to-date understanding of the relevant studies published during the past 10 years (2010-2020). Furthermore, what this review adds is a tighter focus on learning outcomes related to the personal dimension, but with a broader perspective of literacy learning in K-16 education. This review is especially propelled by new literacy studies (e.g., Gee, 2003; Kress, 2003; Mills, 2010), and thus is concerned with the emerging forms of literacy and the interdisciplinary learning potential beyond (traditional) learning of coding. More importantly, this review seeks to address the research gap identified by Kafai and Burke (2015), stating that more documentation on possible challenges pertaining to student game design is needed in the literature. With these thoughts, the present review is therefore conducted to offer a more balanced understanding by attending to some contextual, positive, and negative aspects of student game design, as specified in the previously stated research questions.

3. Method

3.1. Search keywords and sources

The search keywords for the present review included ("game design" OR "game construction" OR "game making" OR "game development") AND (learning OR learners OR students). They were developed according to the purpose of this study, with reference to the previously discussed reviews. The keywords were searched for in titles, abstracts, and author-specified keywords as a preliminary to locating potential articles from a large body of literature in a set of prescribed sources, as specified below.

Three sources of data for the literature search were involved in this review, including electronic databases, journal archives, and reference lists of relevant literature. The methodological decision to go for these search sources was made by consulting relevant publications on guidance for undertaking systematic reviews (Horsley, Dingwall, & Sampson, 2011; Petticrew & Roberts, 2008).

In the digital era, it is commonly believed that searching electronic databases is the most efficient approach to collect data for review studies. Because *ScienceDirect* is one of the largest and most heavily used databases in Taiwan (Ke, Kwakkelaar, Tai, & Chen, 2002), where the authors conducted this study, it was selected as the primary search source for data retrieval.

With an understanding that not all journals are covered by *ScienceDirect*, several refereed journals were also searched. These included: *British Journal of Educational Technology*, *Educational Technology Research and Development*, and *Educational Technology & Society*. They were collectively utilized as the secondary search source due to their reputation as leading journals in the field of education and educational technology, and also for the reason that prior reviews on digital game-based learning (e.g., Hwang & Wu, 2012) have chosen these journals to form their datasets.

Checking reference lists of relevant literature is another avenue to increase the yield of data in review studies, as exemplified by Kafai and Burke's (2015) research synthesis on student game design. Therefore, the reviewed studies cited in the references of the aforementioned reviews (i.e., Kafai & Burke, 2015; Kordaki & Gousiou, 2016; Kordaki & Gousiou, 2017; Li & Tsai, 2013) were manually searched in a snowballing manner as a supplement to the other two search sources of this review.

3.2. Inclusion criteria

Five inclusion criteria were applied during full-text reading of potentially relevant articles to further determine the relevance of a reported study to the present review.

- The study was published during the review period of January 2010 to April 2020.
- The study was reported in a peer-reviewed journal with the Social Science Citation Index (SSCI).
- The study had to be presented as a full-length research article with a robust methodology.
- The study involved K-16 students as the primary participants or game designers.
- The study focused on the use of the game-making approach to facilitate students' literacy learning in some way.

Common examples of excluded articles were those not published during the designated period, those not reported in SSCI journals, those without clear indications of well-designed empirical studies, those focusing on game design by teacher learners or professional game developers rather than school learners, and those addressing other approaches of game use, such as student learning through digital gameplay.

3.3. Coding categories

The coding category of student game designers' educational levels documented the participating students' grade levels based on the K-16 educational system. This was divided into four sub-categories: kindergarten, primary school (grades 1~6), secondary school (grades 7~12), and tertiary or higher education (grades 13~16). A sub-category of mixed was used for studies that recruited participants with different educational levels across settings.

The coding category of game design tools referred to the specific authoring technologies through which the participating students created their own digital games in the reviewed studies. This was not prescribed but allowed for bottom-up emergence in the reviewed studies. A total of 16 game design tools were observed. For those studies without a clear indication of game-making tools, a separate sub-category of unspecified was applied.

The coding category of literacy forms were open coded given the diverse focus of literacy research and the multifaceted nature of literacy. A total of five sub-categories were identified in this review, including (1) basic literacy, (2) intermediate literacy, (3) advanced or disciplinary literacy, (4) 21st century literacy, and (5) new media literacy with respect to digital game design.

The first three sub-categories of literacy forms reflected the traditional understanding of literacy development within disciplines (Shanahan & Shanahan, 2008). Basic literacy, typically acquired in early childhood, referred to the most fundamental skills for being literate in a language, such as reading, writing, and numeracy. Intermediate literacy was defined as the more complex cognitive skills beyond the basic level, which involved domain-specific developmental abilities (e.g., computational thinking in computer science) or domain-general abilities (e.g., analyzing, evaluating, and deep learning strategies). Advanced or disciplinary literacy was considered as specialized knowledge and skills in various subjects or content areas, such as mathematics and science.

The last two sub-categories of literacy forms reflected the contemporary understanding of literacy learning. The so-called 21st century literacy referred to a set of higher-order thinking skills that could be learned and applied across disciplines (Boltz, Henriksen, Mishra, & Deep-Play Research Group, 2015; Conklin, 2011; Trilling & Fadel, 2009). These included, but were not limited to, problem solving, perspective taking, creative thinking, and critical thinking skills. Another relatively new form of literacy that emerged in this review was new media literacy, or more specifically, game design literacy. It was viewed as the ability to properly use and design digital games to express themselves and make meaning out of their learning experiences (Buckingham & Burn, 2007).

As for the coding category of literacy learning orientation, a distinction was made between monodisciplinary and interdisciplinary to understand whether multiple specialized branches of knowledge and skills were embodied in literacy learning and development (Ashby & Exter, 2019). The former referred to a literacy learning orientation that centered on the acquisition of a single branch of knowledge and skills within its disciplinary tradition. An example is teaching students to program a game as a means of helping them develop the targeted computational thinking and programming skills in a computer science course. The latter was an orientation of literacy learning that involved more than one branch of knowledge and skills across traditional disciplinary boundaries. For example, students in a

game design course may design a content-based digital game for history learning, through which to develop their disciplinary literacies in history and computer science.

In answering the last research question, the reviewed studies were initially coded according to whether or not they reported students' perceived challenges when learning with the game-making approach. Details of this category were then inductively coded and analyzed using a thematic analysis approach (Nowell, Norris, White, & Moules, 2017) in order to identify major themes of interest that emerged from the students' point of view. As it turned out, five sub-categories pertaining to the major themes of student perceived challenges were formed.

3.4. Triangulation of literature selection methods

To enhance the research credibility, this study generally followed Petticrew and Roberts' (2008) guidelines for conducting systematic literature reviews in social science, and used multiple methods for data triangulation. First, the keyword-based selection method was adopted to obtain potential articles from the three major sources of data (described in Section 3.1), and 371 articles were initially identified. Next, the criterion-based selection method was utilized to screen the full-texts of all the potentially relevant articles against the five inclusion criteria (described in Section 3.2), and 52 of them remained. Last, the coding-based selection method was employed to assess the eligibility of the shortlisted articles. Two researchers (co-authors of the study) performed this task individually to content analyze each article by applying the coding categories (described in Section 3.3). The inter-coder reliability of the initial analytical results was high (85%). Any coding conflicts were resolved by involving a third researcher through discussion to reach consensus among the research team. Eventually, a final sum of 30 articles were systematically reviewed and reported in this work.

4. Results and discussion

The 30 studies on learning by game-making were included as the review sample, involving a combined total of 2,366 student participants (see Table 1). It was observed that these studies were distributed across various journals, with *Computers & Education* (n = 13) being the most common publication venue, followed by *Educational Technology Research and Development* (n = 7).

4.1. Contextual features of the game-making approach

Table 2 outlines the two contextual features of the game-making approach analyzed in this review. The analytical results pertaining to learners' educational levels indicated that this approach was most frequently applied in secondary education (n = 14), followed by elementary education (n = 6). As expected, this approach was least used with kindergarten students (n = 1) due to its prerequisite of access and ability to learn with technology. Among the five studies with mixed learner groups from different educational levels, one study (Strawhacker & Bers, 2018) recruited children in kindergarten through second grade as participants, and the others (Bossavit & Parsons, 2018; Çakır, Gass, Foster, & Lee, 2017; Papavlasopoulou, Giannakos, & Jaccheri, 2019; Reynolds & Caperton, 2011) recruited participants mostly from middle/high schools. Taken together, most of the reviewed studies involved secondary school learners. Such a finding is contradictory to previous reviews (Kordaki & Gousiou, 2016; Kordaki & Gousiou, 2017) that suggested the frequent use of game-making approach in tertiary or higher education contexts. This is probably due to the difference of years included in this review (2010-2020) and in the previous reviews (2003-2013). It can be inferred that, under the overarching influence of educational digitalization, the game-making approach has been gradually reaching out to younger populations in recent years compared with in the past.

Based on the evidence obtained in this review, teaching and learning with the game-making approach appears more suitable for secondary school learners (and beyond), as they have mastered fundamental literacy skills before moving onto the complex tasks of digital game creation (Kafai & Burke, 2016; Moje, 2015). This finding suggest that when working with younger learners, such as elementary and even kindergarten students, teachers may consider simplifying the technology component in game design so as to lessen students' cognitive load.

Researchers' selection of game-making tools was very diverse, given the 16 different tools identified across the 30 reviewed studies. The most popular tools among the pool were Scratch (n = 6) and Kodu (n = 5), followed by high-end game development engines, including RPG Maker (n = 2), Flash (n = 2), and Neverwinter Nights (n = 2). Other game-making tools were comparatively less popular (n = 1 for each). It further appears that a game design tool is more likely to be selected over others if: (1) it is made available free of charge, and even specifically designed for instructional purposes, as in the cases of Scratch by MIT Media Lab (Ke, 2014) and Kodu by Microsoft (Akcaoglu, 2014); (2) it supports object-oriented visual programming techniques, which is considered more friendly and intuitive for novice learners/programmers (Topalli & Cagiltay, 2018); and (3) it allows for 3D creations, which is deemed more appealing to students (Akcaoglu & Green, 2018).

Among the wide array of the game design tools observed in this review, Scratch and Kodu are comparatively more affordable technologies for learners across different age groups. It is because these two tools are freely accessible, visually appealing, and easy to use. As a result, Scratch and Kodu are suggested as good starting points for educators and researchers interested in the game-making approach, particularly when working with learners with limited or no programming background.

Table 1. List of the reviewed studies and their sample sizes

Study ID	Reviewed studies	Number of participants
S01	Akcaoglu (2014)	18
S02	Akcaoglu and Green (2019)	35
S03	Akcaoglu and Koehler (2014)	44
S04	Allsop (2016)	30
S05	An (2016)	12
S06	Bossavit and Parsons (2018)	6
S07	Çakır et al. (2017)	21
S08	Carbonaro, Szafron, Cutumisu, and Schaeffer (2010)	50
S09	Denner et al. (2012)	59
S10	Dishon and Kafai (2020)	16
S11	Feng and Chen (2014)	232
S12	Gallagher and Grimm (2018)	53
S13	Hava, Guyer, and Cakir (2020)	15
S14	Howland and Good (2015)	55
S15	Hwang, Hung, and Chen (2014)	167
S16	Kalmpourtzis (2019)	34
S17	Kao, Chiang, and Sun (2017)	126
S18	Ke (2014)	64
S19	KovačEvić, Minović, Milovanović, de Pablos, and StarčEvić (2013)	125
S20	Molins-Ruano et al. (2014)	80
S21	Navarrete (2013)	12
S22	Øygardslia and Aarsand (2018)	9
S23	Papavlasopoulou et al. (2019)	157
S24	Reynolds and Caperton (2011)	199
S25	Robertson (2012)	25
S26	Ruggiero and Green (2017)	11
S27	Strawhacker and Bers (2019)	57
S28	Topalli and Cagiltay (2018)	322
S29	Vos, van der Meijden, and Denessen (2011)	235
S30	Yang and Chang (2013)	67
Total	30	2,366

Table 2. Student game designers' adopted tools and educational levels identified in the reviewed studies

Study ID	Game design tools	Student game designers' educational levels				
		Kindergarten	Elementary	Secondary	Tertiary	Mixed
S01	Kodu			X		
S02	Kodu			X		
S03	Kodu			X		
S04	Alice		X			
S05	Gamestar Mechanic			X		
S06	Microsoft Kinet					X
S07	Unity					X
S08	Neverwinter Nights			X		
S09	Stagecast Creator			X		
S10	Scratch & Makey Makey			X		
S11	Scratch		X			
S12	Portal				X	
S13	Kodu			X		
S14	Flip programming language			X		
S15	Kodu		X			
S16	Adobe AIR	X				
S17	Crayon Physics Deluxe			X		
S18	Scratch			X		
S19	Unspecified				X	
S20	C programming language				X	
S21	Flash			X		
S22	RPG Maker		X			
S23	Scratch					X
S24	Flash					X
S25	Neverwinter Nights		X			
S26	Unspecified			X		
S27	Scratch					X
S28	Scratch				X	
S29	Memory Spelen		X			
S30	RPG Maker			X		
Total	16	1	6	14	4	5

4.2. Literacy learning opportunities offered by the game-making approach

Table 3 displays the associated opportunities for literacy development and learning resulting from the use of the game-making approach. A glance at the literacy learning orientations makes it clear that this approach is more achievable as a monodisciplinary (n = 20) than interdisciplinary (n = 10) literacy practice, with computer science being the core disciplinary literacy. Among those studies conducted in monodisciplinary contexts, the development and learning of computer science literacy is generally targeted at the advanced level. In previous reviews (Kordaki & Gousiou, 2016; Kordaki & Gousiou, 2017), similar frequency patterns were observed. It was actually not surprising to find that the development of disciplinary literacy in computer science was the most common literacy learning opportunity available to students due to the nature of digital game design concerning computer skills and domain-specific knowledge of basic programming concepts. One typical example was the study by Howland and Good (2015), in which 55 secondary school students spent eight weeks learning to design their own 3D role-playing games using a simplified programming language, Flip. Comparison of the individual students' pre- and post-tests was used to determine their development with respect to programming knowledge and skills. The results showed that learning-by-game-design was capable of significantly improving the students' disciplinary literacy in computer science.

Delving into the progression of literacy development within disciplines, several studies investigated students' intermediate level of thinking and learning skills (n = 4). For instance, the game-making approach was found to facilitate the students' domain-specific abstraction and reading skills in computer science (Carbonaro et al., 2010; Strawhacker & Bers, 2019) and various domain-general thinking skills, such as organizing, evaluating, and deep

learning skills (e.g., Allsop, 2016; Vos et al., 2011). Only one study (Kalmpourtzis, 2019) applied the game-making approach through the expert-guided use of Adobe AIR in combination with low-tech prototypes to improve kindergarten students' basic level literacy, in this case pre-algebraic patterning.

The review results also revealed the interdisciplinary learning potential of the game-making approach. One-third of the 30 reviewed studies (n = 10) were classified as literacy research with an interdisciplinary learning orientation because they integrated literacy practices of computer science and another discipline. These included two studies each in physics (Gallagher & Grimm, 2018; Kao et al., 2017) and history (Molins-Ruano et al., 2014; Øygardslia & Aarsand, 2018) as well as one study each in mathematics (Ke, 2014), biology (Yang & Chang, 2013), geography (Bossavit & Parsons, 2018), science (Hwang et al., 2014), social studies (An, 2016), and foreign language (Vos et al., 2011). Taking An's (2016) study as an example, seventh graders were engaged to incorporate social studies content into their game design projects using Gamestar Mechanic. The students commented that this interdisciplinary learning experience helped them review what they had previously learned in their social studies class (as design content) through hands-on realization of computer literacy in the form of digital games (as design artifacts). These research instances generally reveal that interdisciplinary learning opportunities enabled by the game-making approach were abundant but selective, as different studies may vary greatly with respect to outcome variables of interest in specific research contexts.

Table 3. Literacy forms and literacy learning orientations identified in the reviewed studies

Study ID	(Traditional) literacy forms			(New) literacy forms		Literacy learning orientation	
	Basic literacy	Intermediate literacy	Advanced literacy	21st century literacy	Game design literacy	Mono-disciplinary	Inter-disciplinary
S01				X		X	
S02				X		X	
S03				X		X	
S04		X				X	
S05			X	X			X
S06			X				X
S07			X			X	
S08		X	X			X	
S09			X			X	
S10				X		X	
S11			X	X		X	
S12			X	X			X
S13				X		X	
S14			X			X	
S15			X	X			X
S16	X			X		X	
S17			X		X		X
S18			X				X
S19			X			X	
S20			X				X
S21				X		X	
S22			X				X
S23			X			X	
S24			X			X	
S25					X	X	
S26				X		X	
S27		X				X	
S28			X			X	
S29		X	X				X
S30			X	X			X
Total	1	4	19	13	2	20	10

In addition to the benefits of cultivating students' traditional literacy development in and across various disciplines, the review results showed that the game-making approach was applicable to developing the learning of so-called new literacy practices, including 21st century literacy (n = 13) and game design literacy (n = 2). For instance, learning by game-making in Yang and Chang's (2013) study was implemented to develop students' 21st century literacy with a focus on critical thinking and their domain-specific academic achievement in biology. In another study that adopted the same approach (Kao et al., 2017), the students' learning outcomes were assessed in terms of physics knowledge acquisition and game design literacy.

Among those studies addressing students' 21st century literacy, problem solving was most frequently examined, with eight of the 13 relevant studies being identified (Akcaoglu, 2014; Akcaoglu & Green, 2019; Akcaoglu & Koehler, 2014; Feng & Chen, 2014; Hava et al., 2020; Hwang et al., 2014; Kalmpourtzis, 2019; Ruggiero & Green, 2017). Other higher-order thinking skills were examined in sporadic studies, including two for creative thinking (Gallagher & Grimm, 2018; Navarrete, 2013), one for critical thinking (Yang & Chang, 2013), one for perspective taking (Dishon & Kafai, 2020), and one for systems thinking and the like (An, 2016). A possible explanation for this finding is that higher-order thinking skills are now gaining increasing attention in literacy education, since such skills are being recognized as essentials for helping students develop into lifelong learners who are competent in dealing with the life demands of the 21st century (Conklin, 2011; Trilling & Fadel, 2009).

While previous reviews have documented the positive effects of the game-making approach on literacy learning in various disciplines (Kafai & Burke, 2015; Kordaki & Gousiou, 2016; Kordaki & Gousiou, 2017; Li & Tsai, 2013), the present review further extends the potential of this approach to the development of 21st century literacy and new media literacy (exemplified by digital game design). This finding opens up new possibilities for literacy educators and researchers to explore various emerging forms of literacy related to the game-making approach. Moreover, educators need to be aware of the interdisciplinary learning potential of the game-making approach, and seek to embed it into a wider curriculum beyond the core discipline of computer science. Doing so may proactively prompt students to ponder the connectivity and interplay between two or more disciplinary literacies, while also nurturing the development of higher-order thinking and other emerging forms of literacy. Creation of content-based digital games is a concrete task that can be executed to achieve the desired outcomes. Following on from this point, it is argued that university students are better candidates than those in lower educational levels. This is largely due to the strong literacy foundation required to perform a complex and satisfactory task as planned (Ashby & Exter, 2019).

4.3. Students' perceived challenges of learning with the game-making approach

Table 4 specifies the studies explicitly reporting learners' accounts of their experience with the game-making approach according to the major themes of challenges which emerged from this review. While learning by making digital games has produced positive findings related to literacy practices in and across disciplines, it may also pose some challenges to participating students, which should not be overlooked. In light of this, each of the reviewed studies was inductively analyzed for students' perceived challenges, if any, based on the presence of relevant dependent variables expressed in the purpose statement and/or research questions. In this review, only a subset of 12 relevant studies out of the entire sample (n = 30) directly examined students' perspectives in this regard. These studies utilized mixed or qualitative methods to generate evidence from diverse data sources, such as interviews, classroom observations, reflection worksheets, open-ended survey questions, and game artifact analyses. As a result of inductive coding and analysis, five major themes were identified, including (1) technological challenges related to the operational use of game design tools, (2) unfamiliarity with game design principles and practices, (3) insufficient time for game design, (4) lack of instructional support during the learning-by-game-making process, and (5) weak or difficult integration of content knowledge into games. Each of these themes is briefly discussed below.

Half of the 12 relevant studies reported that many of the participating students encountered technological challenges as they created games using the designated tools (e.g., Navarrete, 2013). Results of a follow-up analysis revealed that such a technologically-oriented difficulty may be caused by, for example, the daunting task of coding in and of itself (Ke, 2014), the complexity of professional-grade game-making tools like Unity (Çakır et al., 2017), and learner differences, particularly children designers or learners who are less competent in computer literacy (Akcaoglu & Green, 2018). These impeding conditions should be taken into consideration so as to adequately select learner-friendly game-making tools in support of the game-making approach.

Another major challenge identified in this review was learners' unfamiliarity with game design tasks, with five of the 12 relevant studies falling into this category. It is generally agreed in these studies that design by itself is not a simple task, and undoubtedly the multiplicity of digital game design makes it even harder for students to manage. Consequently, assisting students in knowing what design is, what games are, and how these two can be conceptualized together is deemed a necessary first step (Reynolds & Caperton, 2011). Such learner training and preparation must be supplemented with hands-on explorations to prime students' systematic application of design ideas, game mechanics, and creative problem-solving techniques for them to effectively tackle unexpected difficulties (Akcaoglu & Green, 2018).

Time limitation was one common problem reported in three of the 12 relevant studies. From the students' perspective, creating digital games was very time consuming, and they were often overwhelmed by the complicated and iterative task of game design (KovačEvić et al., 2013). To eliminate this problem, enhancing students' time management skills may help them realize and implement their game design plans more efficiently. On the research side, it is recommended to apply the game-making approach in studies with longer durations, rather than one-shot or short-term investigations (lasting from hours to a few weeks).

Table 4. Major themes of student game designers' perceived challenges identified in the reviewed studies

Study ID	Explicit report: Yes/No	The five major themes				
		Technological challenges	Game design difficulties	Time limitations	Lack of support	Weak content integration
S01	No					
S02	Yes	X	X			
S03	No					
S04	No					
S05	Yes		X			X
S06	No					
S07	Yes	X		X		
S08	No					
S09	Yes				X	
S10	Yes					X
S11	No					
S12	No					
S13	Yes		X			
S14	No					
S15	No					
S16	No					
S17	No					
S18	Yes	X				X
S19	Yes			X	X	
S20	No					
S21	Yes	X				
S22	No					
S23	Yes	X				
S24	Yes	X	X	X	X	
S25	Yes		X			
S26	No					
S27	No					
S28	No					
S29	No					
S30	No					
Total	12	6	5	3	3	3

Three out of the subset of 12 relevant studies recognized students' need for guidance in the learning-by-game-making process as a priority area for improvement, particularly when adopting the game-making approach with those who had little or no experience in digital game design. Denner et al. (2012) found that novice game designers were

less likely to persist in the face of setbacks and challenges, and hence extensive instructional support was needed. On this point, some researchers (KovačEvić et al., 2013; Reynolds & Caperton, 2011) have suggested personal consultations and even co-designing with experts as possible solutions to overcome students' unfamiliarity with and uncertainty about what learning-by-game-making might actually entail.

As previously presented, only 10 out of the 30 reviewed studies were implemented with an interdisciplinary learning orientation, and even fewer (n = 3) explicitly examined students' difficulties as they learned with the game-making approach. In such studies, many participating students reported that connecting content knowledge to game design was the most challenging part of the learning experience. As shown in the cases of math game-making in Ke (2014) and historical game-making in An (2016), the students often felt perplexed about how to integrate two disciplines of specialized knowledge and skills in meaningful ways. According to Ke (2014), one strategy to avoid this situation is to provide focused training of content-specific design thinking so as to better prepare student game designers for this integrated intellectual challenge.

All in all, it is evident that digital game creation provides rich and ample literacy learning opportunities, while also bringing some challenges, for students to develop into competent literacy learners who are capable of meeting the societal expectations in today's increasingly competitive environment. Therefore, when implementing the game-making approach for literacy learning in and across disciplines, careful attention should be paid to reduce the common constraints identified in this review.

5. Conclusion

The present review has sought to contribute to the literature by spearheading the game-making approach that is beginning to flourish as a literacy practice in K-16 education. Encouragingly, learning by making digital games has been shown as a promising approach. Prominent reasons are that it is theoretically grounded in constructionist learning, empirically supported by the reviewed studies here, and practically in line with the digitalization of education in contemporary times. On the whole, the findings and implications derived from the present review are anticipated to shed light on the refinement of student game design in future practices.

As with all literature reviews, the sample of this study was limited by the use of search terms, search sources, and search methods for literature collection. The selection of relevant articles was further restricted to those published in SSCI journals during the past 10 years. Many potentially relevant works, particularly "grey literature" (e.g., unpublished dissertations and conference proceedings), were thus excluded from consideration. To complement the focus of this review, meta-analyses that synthesize both published and unpublished studies with a quantitative approach are especially needed to determine the effectiveness of student game design.

Acknowledgement

This study is sponsored in part by the Ministry of Science and Technology in Taiwan under the contract numbers MOST 106-2628-S-327-001-MY3, MOST 106-2628-S-992-303-MY3, and MOST 108-2511-H-008-010-MY3.

References

(References marked with an asterisk indicate studies reviewed in this work.)

*Akcaoglu, M. (2014). Learning problem-solving through making games at the game design and learning summer program. *Educational Technology Research and Development, 62*(5), 583-600.

*Akcaoglu, M., & Green, L. S. (2019). Teaching systems thinking through game design. *Educational Technology Research and Development, 67*(1), 1-19.

*Akcaoglu, M., & Koehler, M. J. (2014). Cognitive outcomes from the Game-Design and Learning (GDL) after-school program. *Computers & Education, 75*, 72-81.

*Allsop, Y. (2016). A Reflective study into children's cognition when making computer games. *British Journal of Educational Technology, 47*(4), 665-679.

*An, Y. J. (2016). A Case study of educational computer game design by middle school students. *Educational Technology Research and Development, 64*(4), 555-571.

Ashby, I., & Exter, M. (2019). Designing for interdisciplinarity in higher education: Considerations for instructional designers. *TechTrends, 63*(2), 202-208.

Boltz, L. O., Henriksen, D., Mishra, P., & Deep-Play Research Group. (2015). Rethinking technology and creativity in the 21st century: Empathy through gaming-perspective taking in a complex world. *TechTrends, 59*(6), 3-8.

*Bossavit, B., & Parsons, S. (2018). Outcomes for design and learning when teenagers with autism codesign a serious game: A Pilot study. *Journal of Computer Assisted Learning, 34*(3), 293-305.

Boyle, E. A., Hainey, T., Connolly, T. M., Gray, G., Earp, J., Ott, M., Lim, T., Ninaus, M., Ribeiro, C., & Pereira, J. (2016). An Update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Computers & Education, 94*, 178-192.

Buckingham, D., & Burn, A. (2007). Game literacy in theory and practice. *Journal of Educational Multimedia and Hypermedia, 16*(3), 323-349.

*Çakır, N. A., Gass, A., Foster, A., & Lee, F. J. (2017). Development of a game-design workshop to promote young girls' interest towards computing through identity exploration. *Computers & Education, 108,* 115-130.

*Carbonaro, M., Szafron, D., Cutumisu, M., & Schaeffer, J. (2010). Computer-game construction: A Gender-neutral attractor to Computing Science. *Computers & Education, 55*(3), 1098-1111.

Chen, M. P., Wong, Y. T., & Wang, L. C. (2014). Effects of type of exploratory strategy and prior knowledge on middle school students' learning of chemical formulas from a 3D role-playing game. *Educational Technology Research and Development, 62*(2), 163-185.

Conklin, W. (2011). *Higher order thinking skills to develop 21st century learners.* Huntington Beach, CA: Shell Education.

Cope, B., & Kalantzis, C. (2000). (Eds.). *Multiliteracies: Literacy learning and the design of social futures.* London, UK: Routledge.

De Freitas, S. (2018). Are games effective learning tools? A Review of educational games. *Educational Technology & Society, 21*(2), 74-84.

*Denner, J., Werner, L., & Ortiz, E. (2012). Computer games created by middle school girls: Can they be used to measure understanding of computer science concepts? *Computers & Education, 58*(1), 240-249.

*Dishon, G., & Kafai, Y. B. (2020). Making more of games: Cultivating perspective-taking through game design. *Computers & Education, 148*, 103810. doi:10.1016/j.compedu.2020.103810

Eseryel, D., Law, V., Ifenthaler, D., Ge, X., & Miller, R. (2014). An Investigation of the interrelationships between motivation, engagement, and complex problem solving in game-based learning. *Educational Technology & Society, 17*(1), 42-53.

*Feng, C. Y., & Chen, M. P. (2014). The Effects of goal specificity and scaffolding on programming performance and self-regulation in game design. *British Journal of Educational Technology, 45*(2), 285-302.

Franciosi, S. J. (2017). The Effect of computer game-based learning on FL vocabulary transferability. *Educational Technology & Society, 20*(1), 123-133.

*Gallagher, D., & Grimm, L. R. (2018). Making an impact: The effects of game making on creativity and spatial processing. *Thinking Skills and Creativity, 28,* 138-149.

Gee, J. (2003). *What video games have to teach us about learning and literacy.* New York, NY: Palgrave Macmillan.

Gilster, P. (1997). *Digital literacy.* New York, NY: Wiley.

*Hava, K., Guyer, T., & Cakir, H. (2020). Gifted students' learning experiences in systematic game development process in after-school activities. *Educational Technology Research and Development*. doi:10.1007/s11423-020-09750-z

Hawlitschek, A., & Joeckel, S. (2017). Increasing the effectiveness of digital educational games: The Effects of a learning instruction on students' learning, motivation and cognitive load. *Computers in Human Behavior, 72,* 79-86.

Horsley, T., Dingwall, O., & Sampson, M. (2011). Checking reference lists to find additional studies for systematic reviews. *Cochrane Database of Systematic Reviews, 8*. doi:10.1002/14651858.MR000026.pub2

*Howland, K. & Good, J. (2015). Learning to communicate computationally with a Flip: A bi-modal programming language for game creation. *Computers & Education, 80*, 224-240.

Hung, H. T., Yang, J. C., Hwang, G. J., Chu, H. C., & Wang, C. C. (2018). A Scoping review of research on digital game-based language learning. *Computers & Education, 126*, 89-104.

*Hwang, G. J., Hung, C. M., & Chen, N. S. (2014). Improving learning achievements, motivations and problem-solving skills through a peer assessment-based game development approach. *Educational Technology Research and Development, 62*(2), 129-145.

Hwang, G. J., & Wu, P. H. (2012). Advancements and trends in digital game-based learning research: A Review of publications in selected journals from 2001 to 2010. *British Journal of Educational Technology, 43*(1), E6-E10.

Kafai, Y. B., & Burke, Q. (2015). Constructionist gaming: Understanding the benefits of making games for learning. *Educational Psychologist, 50*(4), 313-334.

Kafai, Y. B., & Burke, Q. (2016). *Connected gaming: What making video games can teach us about learning and literacy.* Cambridge, MA: MIT Press.

Kafai, Y. B., & Resnick, M. (2012). (Eds.). *Constructionism in practice: Designing, thinking, and learning in a digital world.* New York, NY: Routledge.

*Kalmpourtzis, G. (2019). Connecting game design with problem posing skills in early childhood. *British Journal of Educational Technology, 50*(2), 846-860.

*Kao, G. Y. M., Chiang, C. H., & Sun, C. T. (2017). Customizing scaffolds for game-based learning in physics: Impacts on knowledge acquisition and game design creativity. *Computers & Education, 113*, 294-312.

*Ke, F. (2014). An Implementation of design-based learning through creating educational computer games: A Case study on mathematics learning during design and computing. *Computers & Education, 73*, 26-39.

Ke, H. R., Kwakkelaar, R., Tai, Y. M., & Chen, L. C. (2002). Exploring behavior of E-journal users in science and technology: Transaction log analysis of Elsevier's ScienceDirect OnSite in Taiwan. *Library & Information Science Research, 24*(3), 265-291.

Kordaki, M., & Gousiou, A. (2016). Computer card games in computer science education: A 10 year review. *Educational Technology & Society, 19*(4), 11-21.

Kordaki, M., & Gousiou, A. (2017). Digital card games in education: A Ten year systematic review. *Computers & Education, 109*, 122-161.

*KovačEvić, I., Minović, M., Milovanović, M., De Pablos, P. O., & StarčEvić, D. (2013). Motivational aspects of different learning contexts: "My mom won't let me play this game…". *Computers in Human Behavior, 29*(2), 354-363.

Kress, G. (2003). *Literacy in the new media age.* London, UK: Routledge.

Lankshear, C., & Knobel, M. (2003). *New literacies: Changing knowledge and classroom learning.* Philadelphia, PA: Open 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.

Mills, K. A. (2010). A Review of the "digital turn" in the new literacy studies. *Review of Educational Research, 80*(2), 246-271.

Moje, E. B. (2015). Doing and teaching disciplinary literacy with adolescent learners: A Social and cultural enterprise. *Harvard Educational Review, 85*(2), 254-278.

*Molins-Ruano, P., Sevilla, C., Santini, S., Haya, P. A., Rodríguez, P., & Sacha, G. M. (2014). Designing videogames to improve students' motivation. *Computers in Human Behavior, 31*, 571-579.

*Navarrete, C. C. (2013). Creative thinking in digital game design and development: A Case study. *Computers & Education, 69*, 320-331.

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods, 16*(1), 1-13.

*Øygardslia, K., & Aarsand, P. (2018). “Move over, I will find Jerusalem”: Artifacts in game design in classrooms. *Learning, Culture and Social Interaction, 19,* 61-73.

*Papavlasopoulou, S., Giannakos, M. N., & Jaccheri, L. (2019). Exploring children’s learning experience in constructionism-based coding activities through design-based research. *Computers in Human Behavior, 99,* 415-427.

Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas.* New York, NY: Basic Books.

Papert, S. (1991). Situating constructionism. In I. Harel & S. Papert (Eds.), *Constructionism* (pp. 1-12). Norwood, NJ: Ablex.

Petticrew, M., & Roberts, H. (2008). *Systematic reviews in the social sciences: A Practical guide*. Malden, MA: Blackwell.

Prensky, M. (2008). Students as designers and creators of educational computer games: Who else? *British Journal of Educational Technology, 39*(6), 1004-1019.

Reynolds, R. (2016). Defining, designing for, and measuring “social constructivist digital literacy” development in learners: A Proposed framework. *Educational Technology Research and Development, 64*(4), 735-762.

*Reynolds, R., & Caperton, I. H. (2011). Contrasts in student engagement, meaning-making, dislikes, and challenges in a discovery-based program of game design learning. *Educational Technology Research and Development, 59*(2), 267-289.

*Robertson, J. (2012). Making games in the classroom: Benefits and gender concerns. *Computers & Education, 59*(2), 385-398.

*Ruggiero, D., & Green, L. (2017). Problem solving through digital game design: A Quantitative content analysis. *Computers in Human Behavior, 73,* 28-37.

Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content-area literacy. *Harvard Educational Review, 78*(1), 40-59.

*Strawhacker, A., & Bers, M. U. (2019). What they learn when they learn coding: investigating cognitive domains and computer programming knowledge in young children. *Educational Technology Research and Development, 67*(3), 541-575.

Street, B. (1998). New literacies in theory and practice: What are the implications for language in education? *Linguistics and Education, 10*(1), 1-24.

*Topalli, D., & Cagiltay, N. E. (2018). Improving programming skills in engineering education through problem-based game projects with Scratch. *Computers & Education, 120,* 64-74.

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times.* San Francisco, CA: Jossey-Bass.

*Vos, N., van der Meijden, H., & Denessen, E. (2011). Effects of constructing versus playing an educational game on student motivation and deep learning strategy use. *Computers & Education, 56*(1), 127-137.

*Yang, Y. T. C., & Chang, C. H. (2013). Empowering students through digital game authorship: Enhancing concentration, critical thinking, and academic achievement. *Computers & Education, 68,* 334-344.

Learning Tennis through Video-based Reflective Learning by Using Motion-Tracking Sensors

Chih-Hung Yu[1,4], Cheng-Chih Wu[1,4], Jye-Shyan Wang[2*], Hou-Yu Chen[2,3] and Yu-Tzu Lin[1,4]

[1]Graduate Institute of Information and Computer Education, National Taiwan Normal University, Taiwan // [2]Department of Physical Education, National Taiwan Normal University, Taiwan // [3]Education Center for Humanities and Social Sciences, National Yang-Ming University, Taiwan // [4]Institute for Research Excellence in Learning Sciences, National Taiwan Normal University, Taiwan // chihung@ntnu.edu.tw // chihwu@ntnu.edu.tw // jyeshyan@ntnu.edu.tw // houyuchen@ym.edu.tw // linyt@ntnu.edu.tw

[*]Corresponding author

ABSTRACT: This study proposed a video-based reflective learning approach using motion-tracking sensors to facilitate the learning of tennis skills in a college physical education class by beginning players. The motion-tracking sensors, synchronized with a smartphone video application, were attached to tennis rackets for collecting the students' shot-data. By observing one's practice videos, students could compare their performance with the instructor's demo videos and reflect on the differences for possible improvement. A quasi-experimental method was conducted on two intact classes of students to investigate the effects of the proposed approach. The results showed that students taught by the proposed approach performed better than the traditional approach, exhibited positive attitudes toward learning, and obtained the essence of key tennis techniques. Future implementation should train students how to interpret the sensor collected shot-data so that students can have richer information for reflection.

Keywords: Sensors, Reflective Learning, Tennis, Physical Education, Video-based Learning

1. Introduction

Tennis is an enjoyable sport and often a favorite choice for students taking sports courses. It is provided as a physical education course on an elective basis for college students in Taiwan. Tennis is, however, a sport which students often find difficult to master because it requires the integration of multiple complicated skills. Most beginners have difficulty in mastering fundamental skills such as forehand and backhand groundstrokes. While performing a stroke, four critical temporal phases are involved -- preparation, backswing, forward stroke as well as the follow-through (Knudson & Elliott, 2004). The integration and application of these elements is often difficult for beginners-- it requires the combination of full-body coordination and proper timing of movements. Beginner group tennis classes ordinarily number between twenty and thirty students. As a result of the limitation of a two-hour class per week, students often do not get sufficient feedback from instructors to master the fundamentals of tennis. Even if feedback is provided, students often have limited opportunities to observe their movements—thus, students are unable to connect the feedback provided by the instructor with how they have performed in the class. Another major factor that has been discussed in existing research of sports learning was gender differences. Gender differences exist in motor skills acquisition, including tennis learning (Krumer, Rosenboim, & Shapir, 2016). Physical limitations and psychological tendency are two main reasons attribute to a gender difference in sports learning (Thomas & Thomas, 1988; Vilhjalmsson & Kristjansdottir, 2003). Physical differences such as body mass index (BMI) and muscular endurance benefit males' success in sports activities and therefore enhance their interests and confidence in sports learning. The difference in psychological tendencies affects students' beliefs and reflection quality in sports learning. A mixed-gender grouping is suggested to improve students' performance in physical education courses.

The importance of reflection has been addressed for both student learning and teacher training in physical education (Hanrahan, Pedro, & Cerin, 2009; Groves & O'Donoghue, 2009; Potdevin et al., 2018; Standal & Moe, 2013). Reflection is the process of an individual recapturing their experience, thinking about it, and assessing it. It is the capacity to apply prior experiences to improve subsequent performances in a goal-directed and effective manner (Zimmerman, 2000). A previous study showed that youth athletes who displayed a frequent use of reflection in their practices might attain more success later in their development (Jonker, Elferink-Gemser, de Roos, & Visscher, 2012). Reflection facilitated learners to become more aware of their strengths and weaknesses and help them compare the expected performance with their movements, thereby improving their sports techniques (Panteli, Tsolakis, Efthimiou, & Smirniotou, 2013). It also helped preservice physical education teachers to link their teaching

experiences with pedagogical theories (Garrett & Wrench, 2008; Standal & Moe, 2013). However, Hanrahan, Pedro, and Cerin's (2009) study on sports learning found that the biggest complaint from students was the requirement to take time out from class to complete the reflection forms, and they suggested future study to complete the forms after the class.

To facilitate sports learners to recall the details of their previous performances, video-recording of the learner's movements during a practice session is necessary. Many studies have used video as viewable feedback to improve students' motor skills (Kretschman, 2017; Liebermann et al., 2002; Palao, Hastie, Cruz, & Ortega, 2015). Through the use of video-feedback, students pay more attention to the details of their performances, and better-applied the teacher's feedback to enhance their learning (Nowels & Hewit, 2018). Video-feedback enabled students to understand their performance and benefit from cognitive intervention techniques— especially when engaging in complicated motor tasks requiring power and coordination (Panteli, Tsolakis, Efthimiou, & Smirniotou, 2013). By observing one's practice video combined with the teacher's verbal feedback, students were better able to make significant improvements in their technique as well as having more high-quality practices in class (Palao, Hastie, Cruz, & Ortega, 2015). Potdevin et al. (2018) explored the impact of video-feedback on skill learning in a school-based physical education class. Their findings showed that providing video-feedback coupled with authentic performance data helped novice students reflect on their practice in class, thereby enhanced their gymnastic skills, self-assessment ability, as well as learning motivation. Yet, there were issues with the video-feedback approach-- it required additional teacher time for video recording, reviewing the video, and providing sufficient feedback on the performance. This usually slowed down the pace of instruction (Nowels & Hewit, 2018). By using wearable technology, the logistics of applying video-feedback can be greatly reduced.

With recent advances in IOT (Internet of Things) technology, wearable or ubiquitous sensors can be used to capture personal physical and psychological data in many fields, such as motor learning, language learning, health management, manufacturing processes, and biometric identification (Arif & Kattan, 2015; Blasco, Chen, Tapiador, & Peris-Lopez, 2016; Jou & Wang, 2015; Pan, 2017). In tennis learning, by analyzing and visualizing the information collected from motion capture devices or sensors could help learners better understand their shortcomings while practicing and prevent possible injuries (Oshita et al., 2019; Sharma et al., 2017). Having students access their personal data would facilitate meaningful reflection (Sobko & Brown, 2019). The shot-data of students while learning tennis can now be easily collected and analyzed by using wearable sensors. Büthe, Blanke, Capkevics, and Tröster (2016) used sensors to design a timing analysis system for tennis players. Martin, Bideau, Delamarche, and Kulpa (2016) employed sensors to collect and analyze kinematic, kinetic, and performance changes during prolonged tennis match play to provide quantified information of serve biomechanics. Many tennis sensors are now commercially available at reasonable prices, such as the Babolat Pop, Zepp Tennis Kit, Sony Smart Tennis Sensor, and the Qlipp Tennis Sensor. Some of the sensors have been shown to accurately measure strokes, shot type, ball speed, and hitting volume such as Sony Smart Tennis Sensor (Myers, Kibler, Axtell, & Uhl, 2019) and Babolat Pop (Raymond, Madar, & Montoye, 2019). These sensors detect and record a player's shots and wirelessly connect to smartphones and tablets to help provide information about a player's performance. Some sensors also allow data for each shot to be played back in synchrony with the corresponding recorded video. The slow-motion playback feature allows one to observe the moment of ball impact and check the students' swings as well as footwork. Students have the advantage of being able to reflect on their performances when sensors collecting shot data are paired with recorded videos.

Previous studies have shown the effects of video-feedback on learning tennis. Zheng's (2013) study of freshman beginner tennis classes indicated that video-feedback helped with the students' readiness to learn, ability to learn independently, and perceived deficiencies or incomplete instructor demonstrations. García-González, Moreno, Moreno, Gil, and Del Villar (2013) showed that the combination of video-feedback and questioning on cognitive expertise helped develop adaptations in long-term memory and improve the tactical knowledge of tennis players. Some conflicting results were reported in earlier studies. Emmen, Wesseling, Bootsma, Whiting, and Van Wieringen (1985) found no clear advantages for novices learning tennis when comparing video-feedback groups with the traditional group. However, they indicated, for the scores on form only (in addition to achievement scores), an almost significant interaction effect in favor of video-feedback. They conjectured that the non-significant effect might be due to the video display only providing knowledge of performance (movement information) but lacking a knowledge of results (information about the outcome of the tennis service). A similar study also found that intermediate tennis players gain no apparent advantage when the tennis service is trained to utilize video-feedback instead of traditional training (Van Wieringen, Emmen, Hoogesteger, Bootsma, & Whiting, 1989). This result might be because the video-

feedback was not provided until all members of the group of subjects, who were trained together, had their services recorded. Thus, an improvement of the trainer-subject ratio could lead to a better performance for the video-feedback group.

Resulting from the limited empirical studies provided in the literature, the effect of video-feedback on learning/teaching tennis is inconclusive. In this study, we proposed a video-based reflective learning approach by using wearable technology (as described in Section 2.1) to assist beginners in learning tennis skills in a physical education class. This approach was intended to address the possible drawbacks of the previous studies. Instead of gaining limited feedback from instructors in a high student-instructor ratio group tennis class, this approach would provide students with quality self-reflective feedback. The tennis sensors attached to racquets could offer personal information about the outcome of the learner's shots synchronized with recorded videos. This study explored the effectiveness of our proposed video-based reflective learning approach. The research questions of this study are as follows:

- Does the video-based reflective learning approach using motion-tracking sensors help students learn tennis techniques?
- What are the students' attitudes toward the video-based reflective learning approach?
- How does the video-based reflective learning approach help students learn tennis techniques?

2. Research methods

In this study, we proposed the video-based reflective learning approach based on the literatures surveyed above. A quasi-experimental study was then conducted in order to investigate the effects of the proposed approach. The approach, participants, research instruments, procedures as well as data collection and analysis are described below.

2.1. The Video-based reflective learning approach

Our approach integrated the application of video-feedback, reflective learning, and wearable technology into the teaching of beginner group tennis in a PE setting. The use of video-feedback facilitated students to observe the details of their performance (Nowels & Hewit, 2018) and benefited from cognitive intervention techniques (Panteli, Tsolakis, Efthimiou, & Smirniotou, 2013) such as reflective learning. Reflection enabled students to learn their strengths and weaknesses and helped improve sports skills (Panteli, Tsolakis, Efthimiou, & Smirniotou, 2013). By providing both student's practice videos and the instructor's demonstration videos, students were able to compare their differences with the instructor and came out with improvement ideas. The motion-tracking tennis sensors synchronized video-feedback with the authentic performance data helped students reflect on their practice (Potdevin et al., 2018). In addition, our reflective activities were arranged online and after class to avoid the problem of occupying class time, as reported in Hanrahan, Pedro, and Cerin's (2009) study. We described the details of our approach below.

A typical two-hour group tennis session ordinarily consists of the following stages: (a) Demonstration-- the instructor explains and demonstrates the skills to be learned in the session, (b) Practice-- students practice the skills and the instructor and TAs provide feedback, (c) Wrap-up-- the instructor concludes the session by pointing out common problems students have and re-demonstrating the skills, and (d) After-class activity-- students do assignment after class, for example, watching the instructor's demonstration video on the Internet and answering questions. The proposed video-based reflective approach differed from the traditional approach in two folds (Figure 1). In the Practice stage, we used tablet computers, which synchronized with the tennis sensors, to film the students' practice videos and collect their shot-data as well. The instructor (and/or the TA) then gave students video-feedback onsite. In the After-class stage, students engaged in the reflective process by viewing videos and answering self-assessment questions on a Moodle system. The implementation of both Practice and After-class stages in this study is as follow.

In the Practice stage, students took turns using a racket attached with a sensor. Each student was filmed for approximately 3 minutes. The shot-data, including ball spin, swing speed, swing type, impact position (such as *sweet spot*) and ball speed, could then be displayed in real-time on a smartphone or tablet via Bluetooth. By using the app of the sensor, the students could check the information of each shot they had practiced (see Figure 2). The instructor and/or TAs could then offer immediate feedback onsite. After the class, the TAs would upload the instructor's

demonstration video (Figure 3) along with each student's practice video onto the Moodle system for later use on the After-class reflective activity.

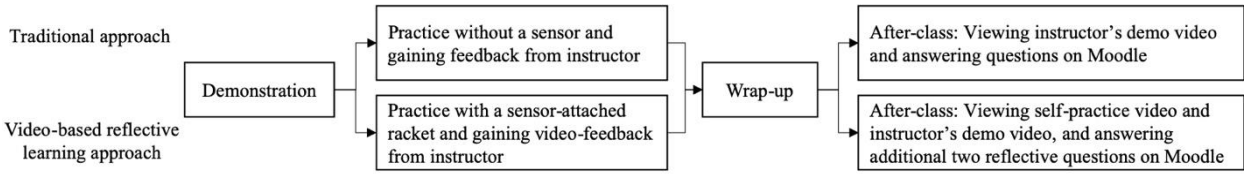


Figure 1. The video-based reflective learning approach vs. the traditional approach

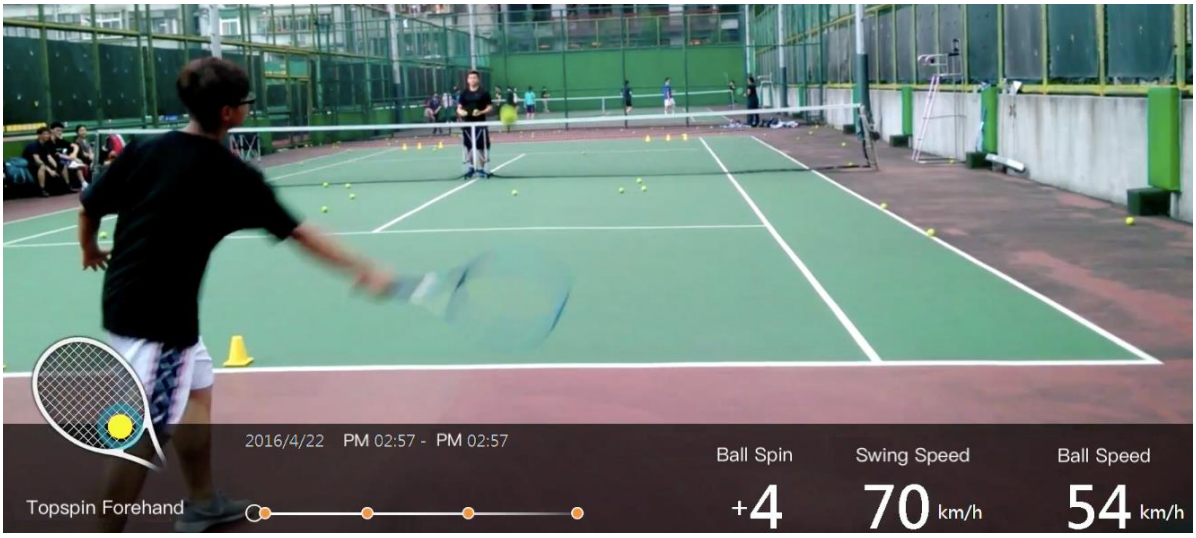


Figure 2. An example of student's shot-data on the practice video

Figure 3. An example of the instructor's shot-data on the demonstration video

At the After-class stage, students were required to engage in a reflection activity on the e-class Moodle system of the university. To place students into the reflective process, students were asked to answer two questions. The first question was, "What skill-related problems have you experienced in class this week?" which provoked students to assess their performance by observing their practice videos. The second question was, "Write down the areas in which you feel that improvements could be made." which enabled students to examine and compare their postures and movement with those of the instructor, as shown in the student's practice video and the instructor's

demonstration videos. For example, a student might find that he or she did not hit the sweet-spot area of the racket (Figure 2, left bottom corner). The student could then compare his or her techniques with those of the instructor (Figure 3, left bottom corner). Students could, additionally, read the shot-data of their swing speeds and ball speeds and find them to be far lower than that of the instructor. A necessary component of the instructional approach is to allow the instructor and TAs to interact with students on the Moodle after the class. Detailed advice about their performances can then be provided to students to improve their skills (see Figure 4). To prevent improper comparisons or judgement among students that might result in incorrect causal attribution (Zimmerman & Campillo, 2003) as well as to keep students' performance private, students were only allowed to view their own reflections, practice videos, and corresponding comments given by the instructor or the TAs.

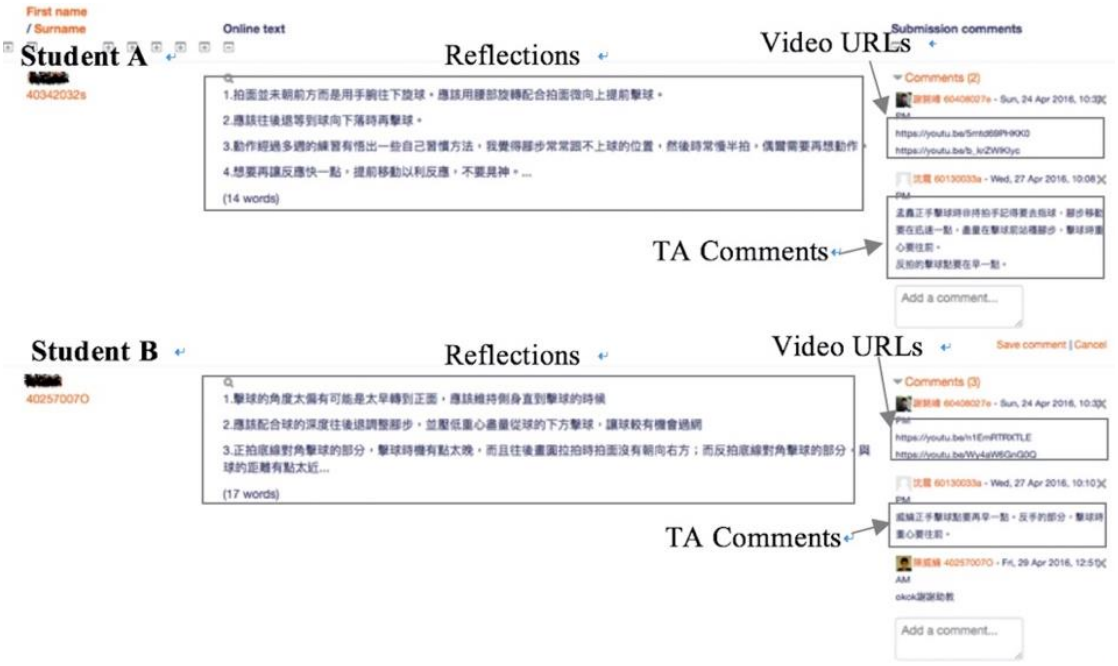


Figure 4. TA's comments on students' reflections on the Moodle

The reflection mechanism in this study (Figure 5) was designed based on Zimmerman and Campillo's self-regulation model (2003). After the practice stage, students were guided for reflection based on their practice videos, instructor's demonstration videos and the corresponding shot-data (video feedbacks). Two reflective questions served as the prompts to facilitate students' self-evaluation and casual attribution process. Students then could get feedbacks from the TAs. After the reflective process, students would be aware of their deficiencies timely and based on which they would develop ideas about posture adjustment for improving their performance. Another learning cycle then restarted from the next practice, ideas generated in the last reflective process helped students focus on their postures and re-exam the effectiveness of their posture adjustment plans.

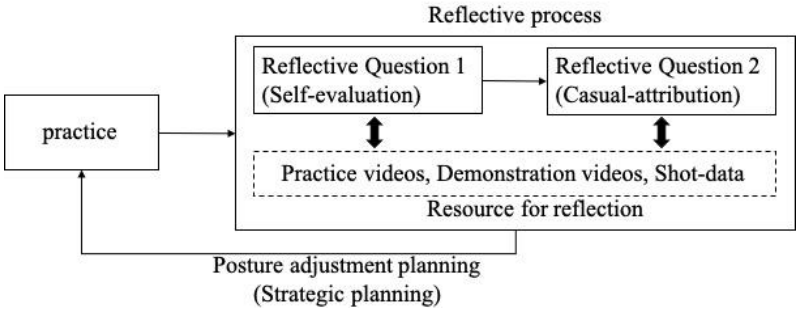


Figure 5. Reflection mechanism in the video-based reflective learning approach

2.2. Participants

The participants in this study were two intact classes of college students who took the PE beginner tennis course at the university. One class with 32 students served as the experimental group and applied video-based reflective activities during instruction. The other class with 30 students served as the control group, applying a traditional approach. After excluding students with high rates of absenteeism, the experimental group consisted of 25 students (10 males and 15 females) and the control group consisted of 25 students (18 males and 7 females). The PE course is required for all students at the university but students can select the sport and level of their choice. Students of the two classes were all beginning players. The instructor explained the purpose, methods, and possible risks of this study to all students at the first class. Students could decide to participate or not, and their decisions would not affect the evaluation of their performance. All participants in this study were over eighteen years old and consented to the research process.

2.3. Research instruments

The research instruments included tennis sensors, Moodle, tennis performance tests, and an attitude questionnaire. We selected the Sony Smart Tennis Sensor for use in our instructional approach. The Smart Sensor, which has been approved by the ITF (International Tennis Federation) for tournament use (ITF, 2014), is an accurate way of measuring hitting data in tennis and can be used by coaches to track player's performance (Myers, Kibler, Axtell, & Uhl, 2019). It equipped with Bluetooth and two different sensors, a 3-axis motion-tracking sensor tracks the movement of the racket, and a vibration sensor acquires data on the strength and the point of impact on the racket head. The sensor, weighing approximately 8 grams, is designed to attach to the grip end of a racket. It is compatible with several tennis racquets; among them, we chose the Head Graphene XT Instinct S racket. Rather than using a smartphone, we used a tablet, the ASUS Transformer Pad TF701T, to support the app of the sensors. The large screen of a tablet allowed for easier viewing during class time so that the instructor could provide students with immediate feedback. In addition, the motion tracking sensors adopted in this study would not collect any individual's biological information or jeopardize the students' health and safety. The Moodle system served as an e-learning platform for the after-class learning activities. Besides delivering instructor's demonstration videos and weekly assignments, Moodle also allowed students in the experimental group to access their practice videos, write reflections, and obtain the corresponding feedback given by the instructor or TAs by using the online texting function (see Figure 4).

Performance tests and an attitude questionnaire were administered to the tennis classes in order to assess the learning outcomes of the students. The performance test was based on the Groundstroke Accuracy Assessment of the International Tennis Number (ITN) scoring standards (ITF, 2004). Four types of skills were assessed in this study, including forehand crosscourt, forehand down the line, backhand crosscourt, and backhand down the line. Five shots were allowed for each of the four skills-- for a total of twenty shots. Each shot was assessed for accuracy, power, and stability. Scores for each shot ranged from zero to seven with a possible maximum score of 140 points. The performance test has clear scoring rubrics based on where the ball lands on the first and second bounce. In addition, each shot of scoring can be correctly judged by the instructor and confirmed with the other two TAs to ensure its reliability.

A ten-item attitude questionnaire (see Appendix) was developed to measure the effects of the proposed video-based reflective learning on students' learning in tennis. Since self-efficacy and intrinsic interest play an important role in sports learning and self-reflection processes (Zimmerman, 2000), these factors were included in the design of the questionnaire to understand students' confidence about their skills (questions 1 & 2), the usefulness of the instructor's demonstration videos (questions 3 & 4), and students' interests in learning tennis (questions 8, 9, & 10). The experimental group was asked three additional questions (questions 5, 6, & 7) related to the video-based reflective activities. Answers to questions used a Likert-type scale, ranging from 1 point (strongly disagree) to 5 points (strongly agree). The questionnaire was reviewed by two tennis educators and two measurement academics to ensure validity. The Cronbach's α values for each dimension ranged from .60 to .83, indicating acceptable reliabilities.

2.4. Procedures

Both classes were held once a week over an 18-week period. Each week had a two-hour class session. All of the course content, teaching schedule, and the instructor and the two TAs were the same for both the control and experimental groups. In each session, both groups of students went through the four instructional stages, as described in Figure 1. The differences between the two groups were that students in the experimental group had video-feedback during the Practice stage. They were able to view their practice videos in addition to the instructor's demonstration videos provided for both groups, and were required to answer two additional questions for the reflective process on Moodle. The control group did not have access to tennis sensors to incorporate the use of practice videos for reflection; however, they were also required to login into the Moodle to answer three to five tactic knowledge questions related to the skills taught during the week, for example, "Which direction should the racquet head face when hitting a ball?" The same questions were also given to the experimental group.

Due to rainy weather conditions and sessions for mid-term and final examinations, students in the experimental group used tennis sensors for a total of 11 weeks. Six of these weeks focused primarily upon forehand and backhand shots and included self-reflective questions. Classes were held on two outdoor tennis courts. On those days where rain prevented outdoor play, students practiced inside a gymnasium or spent time watching instructional or professional tournament videos. At the last lecture of this course, both groups of students were asked to take the performance test and complete the attitude questionnaire. This study did not conduct a pre-test to assess students' performance because all the participants were novice tennis players.

2.5. Data collection and analysis

The data collected for this study were the students' tennis performance test scores, answers to attitude questionnaires, and answers to the instructor's questions on the Moodle. The ANOVA test was conducted to test the performance difference between the two groups. Gender differences have been previously shown to have some effects on student performance in sports (Krumer, Rosenboim, & Shapir, 2016; Thomas & Thomas, 1988; Vilhjalmsson & Kristjansdottir, 2003). As the experimental and control groups had unequal numbers of males and females, a two-way ANOVA was used to account for gender as a possible factor in affecting student performance. The ANOVA test was then conducted to investigate how different learning approaches affect the students' attitudes toward learning. To explore how and why the reflective activities may have affected the students' learning, the answers to the reflective questions provided by the students on Moodle were analyzed and also served as supporting evidence to explain the statistical results.

3. Results and discussions

3.1. Learning performance

Descriptive statistics regarding the student's performance test scores are presented in Table 1. The data showed that students in both groups (control and experimental) averaged somewhere between 20 and 40 points on their performance tests. This score is significantly below the maximum score of 140 points but is common for beginners. On average, students missed half of the 20 balls in the test (the experimental group missed nine balls, while the control group missed 10.2 balls). For each successful shot, students often scored four or lower points (out of seven) due to a lack of ball power.

The results of the ANOVA analysis regarding student performance are presented in Table 2. Partial Eta Squared ($\eta p2$) is presented as a measure of effect size. A $\eta p2$ value between .01 and .06 is classified as a small effect, between .06 and .14 as a medium effect, and .14 or higher as a large effect (Warner, 2012). The analysis showed no significant interactive effects between the gender and the group factors ($F(1, 46) = 3.90$, $p = .05$, $\eta p2 = .08$). Significant differences with a medium effect were observed between experimental and control groups, as well as observed differences between male and female performance. The experimental group performed better than the control group ($F(1, 46) = 6.35$, $p < .05$, $\eta p2 = .12$) and the male students outperformed the female students ($F(1, 46) = 5.08$, $p < .05$, $\eta p2 = .10$) on the performance test. The results suggest that the video-based reflective learning

approach was effective. The result of the gender difference is in accordance with previous studies (Thomas & Thomas, 1988; Vilhjalmsson & Kristjansdottir, 2003).

Table 1. Descriptive statistics for students' performance assessment

	N	*M*	*SD*
Group			
Experimental	25	39.80	12.70
Control	25	34.20	12.54
Gender			
Male	28	39.14	11.58
Female	22	34.27	14.02
Group X Gender			
Experimental/Male	10	40.40	12.76
Experimental/Female	15	39.40	13.09
Control/Male	18	38.44	11.19
Control/Female	7	23.29	9.01

Table 2. The ANOVA results on students' performance assessment

Source	*SS*	*df*	*MS*	*F*	*p*
Group	894.38	1	894.38	6.35*	.015
Gender	715.20	1	715.20	5.08*	.029
Interaction	549.11	1	549.11	3.90	.054
Error	6481.87	46	140.91		

Note. $^{*}p < .05$.

3.2. Attitudes toward the learning activities

Table 3 shows the ANOVA statistical results on student perceptions of the learning activities. Both groups of students showed a medium level confidence on their skills, with a mean around 3.5; and high interest in learning tennis, with a mean above 4.5. There was no difference between the two groups on their confidence in skills ($F(1, 56) = .87$, $p = .35$, $\eta p2 = .02$) and interest in learning ($F(1, 56) = 2.76$, $p = .10$, $\eta p2 = .05$). Although students in both groups showed an interest in enrolling in future tennis classes (question 10), the score of the control group (M = 4.70) was higher than the experimental group (M = 4.39). We speculated that students in the experimental group were less inclined to enroll in tennis courses in the future due to the fact that the video-based reflective approach required additional time and effort as found in the previous studies (Hanrahan, Pedro, & Cerin, 2009; Nowels & Hewit, 2018). They had a comparatively heavier homework load for answering self-reflection questions, and spent more time accommodating the use of sensors in class, as compared to the control group.

As to the effects of videos, both groups perceived the benefit of the instructor's demonstration video, the experimental group showed significant positive attitudes than the control group ($F(1, 56) = 5.31$, $p < .05$, $\eta p2 = .09$). This difference may be because students in the control group were only provided with the instructor's demonstration videos. They were unable to view videos of themselves practicing, which could then act as a basis for comparison with the instructor's demonstration videos. Students in the experimental group, on the other hand, were required to answer self-reflection questions, and were more inclined to view both their own and the instructor's demonstration videos. Students in the experimental group, consequently, were shown to value the instructor's demo videos more highly than the control group.

In addition to the results on Table 3, students in the experimental group were asked additional three questions about the reflective activities. The overall satisfaction level of the students was high, being over four points out of a maximum of five. Students agreed using techniques such as watching their own practice videos (question 5, M = 4.32), corresponding shot-data (question 6, M = 4.16), and TA feedback (question 7, M = 4.16) would be helpful in learning tennis skills. Note that the numbers of students (N) in Table 3 are slightly different from those in Table 1. It is because the attitude questionnaire was conducted at the end of course survey and was filled out anonymously, thus the statistical analysis could not exclude students with high rates of absenteeism.

Table 3. The ANOVA results on students' attitudes toward the learning activities

Dimension	Group	*N*	*M*	*SD*	*F*	*p*
Confidence in skill	Control	27	3.59	.83	.87	.35
	Experimental	31	3.40	.71		
Video effects	Control	27	4.26	.54	5.31*	.03
	Experimental	31	4.58	.52		
Interest	Control	27	4.75	.38	2.76	.10
	Experimental	31	4.53	.58		

Note. $^{*}p < .05$.

3.3 Students' reflection on the Moodle

To explore what kinds of technique problems the students reflected on, six weeks of student answers to reflective questions on Moodle, which focused on the forehand and backhand shots, were collected and analyzed. A total of 135 posts were analyzed by two raters (the first and the third authors) based on the six critical tennis techniques taught in class. Both of the raters reviewed all the posts and discussed the discrepancies between their analysis to reach agreement. Since each post might cover more than one technique problem, we counted the number of times that each key technique was mentioned in the answers by students and calculated a total of 248 mentions. The percentages of each technique mentioned are presented in Table 4. Overall, the most significant problem that students encountered was related to their tennis form. Students felt that the major reasons for making poor shots stemmed from their inability to control their wrist and/or racket angle (technique 1, 24.19%) as well as an uncompleted swing path (technique 2, 24.19%). How to execute the proper timing when hitting a ball (technique 3, 23.79%) and maintaining balance of body (technique 4, 20.56) were also common challenges for tennis beginners. It was found that most of the reflections were key tennis skills that the instructor addressed in the class, only about 6 percent of posts related to other issues (technique 7) such as control of swing speed, their expectations for improvement, or feels about their performance. This demonstrated that the proposed approach promotes student engagement in meaningful reflection and active thought. By examining their own techniques, which were subsequently compared to the instructor's demonstration video, students were able to discover flaws in their skills and devise methods for improvement.

On the other hand, less than one percent of the reflections mentioned problems related to hitting the ball on the sweet spot (technique 5) or identifying the flight trajectory of an approaching ball (technique 6). The technique of hitting on the sweet spot involves many issues such as the timing or the racket angle for making a shot. Although the sweet spot information (and other shot data) was displayed on the practice videos, students often had problems to interpret the data and connect it to their form and movement; thus, resulting in very few reflections. For few reflections on the flight trajectory of an approaching ball, we found it was because that our camera was placed right behind the players--- which prevented us from video-taping the proper angle of the flight path of a ball (see Figure 2 and 3). Future implementation should address these problems.

Table 4. Percentages of students' reflection on the six key tennis techniques ($N = 248$)

Key technique		*N*	%
1.	Control of the angle of the wrist and racket when hitting the ball	60	24.19
2.	Whether the swing path is proper	60	24.19
3.	Proper timing for making a shot	59	23.79
4.	Maintaining balance and shifting the center of gravity of body	51	20.56
5.	Whether hitting the ball on the sweet spot	2	0.81
6.	Identifying the flight trajectory of an approaching ball	1	0.40
7.	Others (swing speed, expecations, emotions)	15	6.05

Below we present several reflective posts from students as supporting evidences on the effects of our approach--from simply observing and self-assessing on one's skills, to reflecting on one's skills and comparing those with the instructor. For example, a student observed that he failed to turn his body and hit the ball too late:

I didn't turn my body aside for a forehand shot, and hit the ball too late for a backhand shot. I have to make use of the strength from turning the body.

Or, one found that he did not hit a ball with proper timing and fail to control the angle of the racket:

I didn't hit the ball at the right time. The racket faced a little too up so that the ball flew too high. My posture is not correct for bringing the ball up for a drop shot. I have to practice more.

The advantages of examining one's performance in the practice video and comparing those with the instructor's (as shown in the demonstration video) were also addressed by the students. For example,

Because I didn't keep my racket facing forward (it faced up), the ball flew unstably. I also have to turn my body more to locate the center of gravity properly. In addition, my steps were not practiced enough. My postures were quite different from the instructor's.

Additionally, some students would compare their shot data (as in Figure 2) with those of the instructor's and conclude ways for improvement. For example,

The spin values of the instructor's hit were all +3, but mine were all +2, therefore I have to turn my body more. The swing speed of the instructor's hit were 57 and 53, respectively, but mine were 47, 48, and 49. I have to firm up my wrist. The ball speed of the instructor's hits were 35 and 33, but mine were 37, 40, and 41. Improper exertion increased the ball speed. The ball speed should be transferred to spinning. I have to practice more to grasp the skills.

The results echo previous study where self-explanation practice prompts the students to recognize links between the knowledge or skills they have learned, and allows them to identify and address gaps in their understanding (Bisra, Liu, Nesbit, Salimi, & Winne, 2018). Students attributed their poor performances to the incorrect hitting postures based on both videos and TA's feedbacks, and planned for the next practice to correct the flaws. Moreover, some of students would compare their shot-data with the instructor's and conclude possible ways to adjust their postures. The reflective approach helped students to correct their hitting postures and improve their performance. However, the level of students' reflection might be a factor contributing to the effects, future studies should be conducted to investigate how students corrected their postures according to the reflections.

4. Issues regarding the use of sensors and videos

One factor that might limit the implementation of the use of sensors in beginner tennis classes is the cost. The ones utilized here cost approximately $200 each. Rackets utilized in the study were somewhat expensive, costing approximately $140 each. For many colleges and universities, equipment use might be expensive to implement fully, although the costs should decrease with increased demand. In the study, since current prices made supplying each student with equipment limited, we decided that students in each subgroup would share the use of sensors. Students used the sensors only during certain parts of the class period and this presented no major problems. It would be better if the sensors could be adaptable to a wider selection of rackets, eliminating the need to purchase new rackets to match various sensors. If individual users could enter their racket parameters, such as weight, length, and racket-face area, into the application, then the use of tennis sensors would be enhanced.

For conducting reflective activities on the Moodle, the students' practice videos had to be transferred from the tablets to a computer, and then uploaded to YouTube. Two potential problems were encountered which were the speed of the transfer and the time commitment. Transferring the videos from the tablets required using dedicated software (Sony PlayMemories) and each video had to be transmitted one-by-one. One component of the study was a requirement for adequate broad bandwidth and storage in order to view and to store videos of both the instructor's demonstration and the students' practice. The solution in the present study was to upload videos to a private YouTube channel, and then post each video's URL on the individual students' Moodle accounts, where files could be viewed by clicking on a link. The use of the YouTube format also allowed students to view their own practice videos in a timely manner. Transferring videos, uploading videos, and posting links all required additional time and effort from the instructor and TAs.

Some additional problems were encountered with the use of a built-in app designed for the sensors in the study. The app is designed for use by single users but not by multiple users. When sensors were shared by students, the app's "summary of all shot-data" read out as a single person's total statistics. This made it difficult to analyze overall

performance from individual students in the group setting. One possible solution would be to install the app on each student's smartphone or tablet computer, then pair the sensor with a student's mobile device during their turn to use the sensor. Initially, this method would have added to the technical difficulty of installing and pairing the app as well as having taken longer to execute.

5. Suggestions for future implementation

In this study, students were able to view their shot-data in the practice videos, but additional time could have been spent on teaching students how to better interpret their shot-data information to improve their tennis form. For example, if students were able to pay special attention to the sweet-spot feedback (Figure 2, left bottom corner) for each shot, it would help them to understand where their problems lay. Students would find that not hitting the ball on the sweet spot usually had to do with their problems on shot timing, firming one's wrist, and racket angle. Furthermore, the sensor's feedback on ball-spin can also shed light on problems with the students' swing trajectories. Students should be taught to use the practice videos to compare their best and worst ball-spins and take steps to correct swing trajectory. Future implementation of using sensors should stress interpreting the shot-data information, which can serve as clues for reflecting on one's practice and comparing the differences with the instructor's form. In addition, filming videos from different shooting angles might help students observe and identify the flight trajectory of an approaching ball. Our study also exhibited a possible drawback indicated by Van Wieringen, and et al. (1989)-- the time span between performing the shot and receiving video-feedback was too long. Future implementation may either increase the TAs or re-schedule the flow of recording video and giving feedback. Reflection is important when learning proper sports techniques, but actual practice is an essential component following the reflection so that the reflected skills can be consolidated. In this study, most students were able to pinpoint their problems when in reflection, but improvements in form were limited because following-up practice sessions were not planned in our approach. Future implementation should place greater emphasis upon immediate practice on the court after students have watched and reflected on their practice videos. The time allotment for each phase may need to be adjusted to incorporate the reflect-then-practice strategy into the instructional approach. However, the level of reflection might also affect students' learning (Kember et al., 2000). Measuring students' reflection level will help investigate more deeply about the relationship between reflection and learning performance. Qualitative research methodologies (e.g., grounded theory research) might also be helpful for exploring implicit effects. Moreover, further study with larger sample size should be conducted to improve the external reliability of the results. Finally, to better control the video effects, students in the control group could also be provided with the self-practice videos so that the effects of videos could be balanced.

6. Conclusion

Learning tennis is often quite difficult because many factors come into play in the course of making a shot. Students are often unable to make proper judgments about running and hitting the ball during the time of play. Through the use of motion-tracking sensors and video-feedback, students could reflect on their own performance and compare those with the instructor's, which in turn helped them form better mental images and concepts about properly hitting a tennis ball. The results of this study showed that the experimental group had better learning outcomes than the control group, and was able to grasp the essence of key tennis skills-- hitting the ball at the appropriate time, controlling the wrist and angle of the racket, maintaining balance when swinging the racket, and whether the swing path is proper. The instructor's demonstration videos were particularly useful for the experimental group in that they were used to compare with the students' video-recorded performances and conduct their reflections. Since physical skills like playing tennis are not easy to be portrayed and explained clearly, the quantized and visualized data from sensors and videos could serve as an objective basis for student observation and comparison.

The use of sensors, videos, and self-reflection in this study was shown to be beneficial. For future implementation, we suggested that students could be trained to interpret the sensor's shot-data; video-feedback could be immediately followed the practice; follow-up practices could be scheduled immediately after reflections are completed, and the use of sensors should avoid interfering with the students' practice times.

Acknowledgement

This research was sponsored by the Ministry of Science and Technology, Taiwan under Grant no. MOST 107-2511-H-003-029-MY3.

References

Arif, M., & Kattan, A. (2015). Physical activities monitoring using wearable acceleration sensors attached to the body. *PloS one, 10*(7), e0130851. doi:10.1371/journal.pone.0130851

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing self-explanation: A Meta-analysis. *Educational Psychology Review, 30(*3), 703-725. doi:10.1007/s10648-018-9434-x

Blasco, J., Chen, T. M., Tapiador, J., & Peris-Lopez, P. (2016). A Survey of wearable biometric recognition systems. *ACM Computing Surveys, 49*(3), 1-35. doi:10.1145/2968215

Büthe, L., Blanke, U., Capkevics, H., & Tröster, G. (2016). A Wearable sensing system for timing analysis in tennis. In *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)* (pp. 43-48), San Francisco, CA: IEEE. doi:10.1109/BSN.2016.7516230

Emmen, H. H., Wesseling, L. G., Bootsma, R. J., Whiting, H. T. A., & Van Wieringen, P. C. W. (1985). The Effect of video-modelling and video-feedback on the learning of the tennis service by novices. *Journal of Sports Sciences, 3*(2), 127-138. doi:10.1080/02640418508729742

García-González, L., Moreno, M. P., Moreno, A., Gil, A., & Del Villar, F. (2013). Effectiveness of a video-feedback and questioning programme to develop cognitive expertise in sport. *PloS one, 8*(12), e82270. doi:10.1371/journal.pone.0082270

Garrett, R., & Wrench, A. (2008). Connections, pedagogy and alternative possibilities in primary physical education. *Sport, Education and Society, 13*(1), 39-60. doi:10.1080/13573320701780514

Groves, M., & O'Donoghue, J. (2009). Reflections of students in their use of asynchronous online seminars. *Educational Technology & Society, 12*(3), 143-149.

Hanrahan, S. J., Pedro, R., & Cerin, E. (2009). Structured self-reflection as a tool to enhance perceived performance and maintain effort in adult recreational salsa dancers. *The Sport Psychologist, 23*(2), 151-169. doi:10.1123/tsp.23.2.151

International Tennis Federation (ITF). (2004). *International tennis number manual*. Retrieved from http://www.tennisplayandstay.com/media/131803/131803.pdf

International Tennis Federation (ITF). (2014). *Player analysis technology approval report-SONY Smart Tennis Sensor* (Serial No. 2001590). Retrieved from https://www.itftennis.com/media/1448/sony-smart-tennis-sensor-report.pdf

Jonker, L., Elferink-Gemser, M. T., de Roos, I. M., & Visscher, C. (2012). The Role of reflection in sport expertise. *The Sport Psychologist, 26*(2), 224-242. doi:10.1123/tsp.26.2.224

Jou, M., & Wang, J. (2015). The Use of ubiquitous sensor technology in evaluating student thought process during practical operations for improving student technical and creative skills. *British Journal of Educational Technology, 46*(4), 818-828. doi:10.1111/bjet.12173

Kember, D., Leung, D. Y., Jones, A., Loke, A. Y., McKay, J., Sinclair, K., Tse H., Webb, C., Wong, F. K. Y., Wong, M., & Yeung, E. (2000). Development of a questionnaire to measure the level of reflective thinking. *Assessment & Evaluation in Higher Education, 25*(4), 381-395. doi:10.1080/713611442

Kretschmann, R. (2017). Employing tablet technology for video feedback in physical education swimming class. *Journal of e-Learning and Knowledge Society, 13*(2), 103-115. Retrieved from https://www.learntechlib.org/p/188114/

Krumer, A., Rosenboim, M., & Shapir, O. M. (2016). Gender, competitiveness, and physical characteristics: Evidence from professional tennis. *Journal of Sports Economics, 17*(3), 234-259. doi:10.1177/1527002514528516

Knudson, D., & Elliott, B. (2004). Biomechanics of tennis strokes. In G. K. Hung, & J. M. Pallis (Eds.), *Biomedical Engineering Principles in Sports* (pp. 153-181). Boston, MA: Springer. doi:10.1136/bjsm.2005.023150

Liebermann, D. G., Katz, L., Hughes, M. D., Bartlett, R. M., McClements, J., & Franks, I. M. (2002). Advances in the application of information technology to sport performance. *Journal of Sports Sciences, 20*(10), 755-769. doi:10.1080/026404102320675611

Martin, C., Bideau, B., Delamarche, P., & Kulpa, R. (2016). Influence of a prolonged tennis match play on serve biomechanics. *PloS one, 11*(8), e0159979. doi:10.1371/journal.pone.0159979

Myers, N., Kibler, W., Axtell, A., & Uhl, T. (2019). The Sony Smart Tennis Sensor accurately measures external workload in junior tennis players. *International Journal of Sports Science & Coaching, 14*(1), 24-31. doi:10.1177/1747954118805278

Nowels, R. G., & Hewit, J. K. (2018). Improved learning in physical education through immediate video feedback. *Strategies, 31*(6), 5-9. doi:10.1080/08924562.2018.1515677

Oshita, M., Inao, T., Ineno, S., Mukai, T., & Kuriyama, S. (2019). Development and evaluation of a self-training system for tennis shots with motion feature assessment and visualization. *The Visual Computer, 35*, 1517-1529. doi:10.1007/s00371-019-01662-1

Palao, J. M., Hastie, P. A., Cruz, P. G., & Ortega, E. (2015). The Impact of video technology on student performance in physical education. *Technology, Pedagogy and Education, 24*(1), 51-63. doi:10.1080/1475939X.2013.813404

Pan, W. F. (2017). The Effects of using the Kinect motion-sensing interactive system to enhance English learning for elementary students. *Educational Technology & Society, 20*(2), 188-200.

Panteli, F., Tsolakis, C., Efthimiou, D., & Smirniotou, A. (2013). Acquisition of the long jump skill, using different learning techniques. *The Sport Psychologist, 27*(1), 40-52. doi:10.1123/tsp.27.1.40

Potdevin, F., Vors, O., Huchez, A., Lamour, M., Davids, K., & Schnitzler, C. (2018). How can video feedback be used in physical education to support novice learning in gymnastics? Effects on motor learning, self-assessment and motivation. *Physical Education and Sport Pedagogy, 23(*6), 559-574. doi:10.1080/17408989.2018.1485138

Raymond, C. J., Madar, T. J., & Montoye, A. H. (2019). Accuracy of the Babolat Pop sensor for assessment of tennis strokes in structured and match play settings. *Journal of Sport and Human Performance, 7*(1). doi:10.12922/jshp.v7i1.146

Sharma, M., Srivastava, R., Anand, A., Prakash, D., & Kaligounder, L. (2017). Wearable motion sensor based phasic analysis of tennis serve for performance feedback. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5945-5949). New Orleans, LA: IEEE. doi:10.1109/ICASSP.2017.7953297

Sobko, T., & Brown, G. (2019). Reflecting on personal data in a health course: Integrating wearable technology and ePortfolio for eHealth. *Australasian Journal of Educational Technology, 35*(3), 55-70. doi:10.14742/ajet.4027

Standal, Ø. F., & Moe, V. F. (2013). Reflective practice in physical education and physical education teacher education: A Review of the literature since 1995. *Quest, 65*(2), 220-240. doi:10.1080/00336297.2013.773530

Thomas, J. R., & Thomas, K. T. (1988). Development of gender differences in physical activity. *Quest, 40*(3), 219-229. doi:10.1080/00336297.1988.10483902

Van Wieringen, P. C. W., Emmen, H. H., Bootsma, R. J., Hoogesteger, M., & Whiting, H. T. A. (1989). The Effect of video-feedback on the learning of the tennis service by intermediate players. *Journal of Sports Sciences, 7*(2), 153-162. doi:10.1080/02640418908729833

Vilhjalmsson, R., & Kristjansdottir, G. (2003). Gender differences in physical activity in older children and adolescents: the central role of organized sport. *Social Science & Medicine, 56*(2), 363-374. doi:10.1016/S0277-9536(02)00042-4

Warner, R. M. (2012). *Applied statistics: from bivariate through multivariate techniques*. Thousand Oaks, CA: Sage Publications.

Zheng, Z. (2013). The Experimental study on computer video feedback in tennis serve teaching effect. In *2013 International Workshop on Computer Science in Sports (IWCSS)* (pp. 262-264). Wuhan, China: Atlantis Press. doi:10.2991/iwcss-13.2013.68

Zimmerman, B. J. (2000). Attaining self-regulation. A Social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 13-39). San Diego, CA: Academic Press. doi:10.1016/B978-012109890-2/50031-7

Zimmerman, B. J., & Campillo, M. (2003). Motivating self-regulated problem solvers. In J. E. Davidson & R. J. Sternberg (Eds.), *The Psychology of Problem Solving* (pp. 233-262). Cambridge, UK: Cambridge University Press.

Appendix: Attitude questionnaire

Dimensions	Questions
Confidence in skills	1. I have confidence with my forehand.
	2. I have confidence with my backhand.
Video effects (and reflective activities)	3. Watching the instructor's demonstration video is helpful.
	4. I like comparing my playing with the instructor's playing on the demonstration video.
	5. Watching my practice video is helpful.*
	6. The shot information (sweet spot, speed, spin, etc.) shown on the video is helpful. *
	7. TAs' feedback on my reflections on Moodle is helpful. *
Interests in learning	8. I like the tennis class.
	9. I became more interested in tennis.
	10. I will continue to enroll in a tennis class.

Note. *Questions for the experimental group only.

Enhancing Post-secondary Writers' Writing Skills with a Chatbot: A Mixed-Method Classroom Study

Michael Pin-Chuan Lin* **and Daniel Chang**
Simon Fraser University, Faculty of Education, Canada // pla85@sfu.ca // dth7@sfu.ca
*Corresponding author

ABSTRACT: In the present study, we developed a chatbot that helps teachers to deliver writing instructions. By working with the chatbot, the post-secondary writers developed a thesis statement for their argumentative essay outlines, and the chatbot helped the writers to refine their peer review feedback. We conducted a preliminary analysis of the effect of a chatbot on these writers' writing achievement. We also collected several student testimonials about their chatbot experiences. Several important pedagogical and research implications for chatbot-guided writing instructions and the use of learning technology have been addressed.

Keywords: Chatbot, Intelligent tutoring system (ITS), Learning tool, Writing skills, Classroom study

1. Introduction

Since the past decades, many educators have started to diversify instructions by adopting educational technology in classrooms. Several intelligent tutoring systems (ITS) have been developed, and these ITS are often built upon specific algorithms that offer learners individualized instructions or evaluate students learning products (Kerly, Hall, & Bull, 2007; Ma, Adesope, Nesbit, & Liu, 2014; Rodríguez-Gil, García-Zubia, Orduña, Villar-Martinez, & López-De-Ipiña, 2019; Vanlehn, 2006). However, Murray (1999) pointed out some limitations of ITS, including low fidelity from student perspectives, limited instructional values, lack of student modelling, and limited interactivity. To overcome these limitations, some studies recommend that if a chatbot is programmed on the supplementary side of ITS, it may help to facilitate a real-time dialogue that supports thinking and learning processes. Nevertheless, some research studies have not distinguished the differences between ITS and a chatbot (Kerly et al., 2007; Wang & Petrina, 2013). Traditionally, an ITS often takes over an instructor's role by presenting learning materials and offering feedback to students (Song, Oh, & Rice, 2017). On the other hand, a chatbot is often a supplementary conversational program that interacts with users synchronously, such as human-like conversations, question answering, user support, or tutoring (Abbasi & Kazi, 2014; Clarizia, Colace, Lombardi, Pascale, & Santaniello, 2018; Kerly et al., 2007; Pereira & Díaz, 2018). Especially, Jain, Kumar, Kota, and Patel (2018) defined chatbots as "text-based, turn-based, task-fulfilling programs, embedded within existing platforms" (p. 904).

In educational research, a chatbot has always been implemented with a specific instructional intention, such as promoting class engagement (Kerly et al., 2007) or promoting critical thinking (Goda, Yamada, Matsukawa, Hata, & Yasunami, 2014). For instance, teachers might use a chatbot to promote critical thinking. Goda et al. (2014) carried out two case studies involving a total of 130 university students, divided into two groups, with each participating in two successive class periods. One class period was experimental, where students conversed with an ELIZA-based chatbot, and another class period was a comparison experience where students listed their thoughts and searched relevant information on the Internet. The results from case study 2 revealed a positive impact of the chatbot on students' awareness of critical thinking and inquiring mindset.

Another specific instructional intention might be promoting language learners' conversational skills. Studies in applied linguistics have found that a chatbot might be developed to improve language learners' conversational skills (Fryer & Carpenter, 2006). Particularly, Fryer and Carpenter (2006) mentioned: "there is yet no chatbot designed from the bottom up to meet the needs of FLL (foreign language learners) students" (p. 12). They have suggested how a learner can freely engage conversations with a chatbot, review what has been talked about in a transcript format, and self-analyze the transcription from the interaction. For teachers, a chatbot might also be useful because the conversation, presented in a transcript format, records a student's progression as well as their needs for future learning. Furthermore, when learners are struggling with the learning materials, a chatbot can also offer just-in-time guidance and solve basic problems for students immediately (Fryer & Carpenter, 2006; Pereira & Díaz, 2018).

Upon careful examination of educational technology literature, we have realized that chatbot-guided writing activities are relatively rare and scarce. Mainly, previous research has established a chatbot-led pre-discussion activity to improve students' critical thinking skills (Goda et al., 2014). In our study, a writing chatbot was introduced as part of a university disciplinary writing class activity, acting as a supplementary activity to in-class writing instructions. Acquiring writing skills is not a linear process, yet it is an interactive social process that requires multiple, multichannel input and output between individuals and the chatbot itself. Theoretically speaking, building a chatbot that assists writing instructions might fulfill novice writers' needs to initiate a dialogue when help is needed right away (Cazden, 1988; Edwards & Mercer, 2013).

The current study further advances Goda's et al. study (2014) in a way that the chatbot can help students to generate their thesis statement for an argumentative essay outline. Through working with the chatbot, students will be able to evaluate their ideas for their thesis statements. Overall, the expected outcome of the chatbot is to assist students with two major events in the writing process: drafting a thesis statement for the essay outline and learning to offer peer feedback on the outlines. Data collected as a result of learners' conversation with the chatbot point to the following essential questions:

- When students work with a chatbot for their thesis statements and peer feedback, do these chatbot-led activities enhance their writing achievement?
- How do students perceive the use of chatbot in a university classroom?

These two questions are crucial in understanding how a chatbot can supplement writing instructions, as instructors are not always available when students need help (Song et al., 2017; Xu & Wang, 2006). Based on the social constructivist theory of learning, a conversation is the key to learning (Kalina & Powell, 2009). If so, engaging students in a chatbot-led conversation might support the writing and learning activities. We thus expect that conversations initiated by the chatbot might support student writers in composing their thesis statements and peer feedback.

2. Literature review

2.1. Design framework of Chatbot

The antecedent of a chatbot is ELIZA, a computerized system capable of parsing human's natural language and initiating conversations. Eliza parsed user input, identified key phrases from its backend template and selected corresponding responses (Weizenbaum, 1996). Failure occurs when user input does not exist in its backend text template (Fryer & Carpenter, 2006; Kirakowski, O'Donnell & Yiu, 2009). However, Eliza's programmed scripts and unnatural interaction are the primary design limitations (Goda et al., 2014). Subsequent chatbot research has suggested some important design features, such as field-specific scope control, use of multimedia resources, and fallback response (Jain et al., 2018).

2.1.1. Field specific scope control

A domain-specific chatbot is the most favourable type of chatbot because it minimizes the unrestricted creativity of human language capacity (e.g., Jain et al., 2018; Luger & Sellen, 2016). Especially, Luger and Sellen (2016) have found that a chatbot is better developed within limited scopes, functions and purposes so that users can better facilitate effective interaction with the chatbot and have better ability to fulfill their purposes. For example, Ghose and Barua (2013) developed an FAQ chatbot using AIML (Artificial Intelligence Markup Language). The purpose of the chatbot was to serve as an undergraduate students' advisor, and it was designed for assisting course and admission information retrieval. The research also evaluated the conversation accuracy rate by investigating the student-to-chatbot interaction logs. Results showed the students were more satisfied with the domain-specific chatbot. Their results imply a domain-specific chatbot is more helpful than a chatbot that understands everything. Moreover, a chatbot (Lucy) developed by Wang and Petrina (2013) failed to recognize student input accurately when Lucy was programmed for multiple areas. To solve this issue, Lucy was redesigned and programmed with more specific domains that only handle topics related to tourism. Taken together, designing a domain-specific chatbot seems to increase the accuracy of chatbot's response.

2.1.2. Use of multimedia resources

Combining text and multimedia facilitates the interaction between a chatbot and a user (Jain et al., 2018). A combination of multimedia resources is effective in promoting engagement, especially if the chatbot has a text-to-speech function in the field of education (Fryer & Carpenter, 2006). Furthermore, if a chatbot can begin the conversation by playing a game with a user, then users might be more engaged and motivated to explore the chatbot (Luger & Sellen, 2016). These findings imply that chatbot-guided instructions might consider integrating several multimedia features for students to engage their attention.

2.1.3. Fallback response

Fallback response is a mechanism triggered when a user intent is out of space that may cause a conversation to fail. Kerly et al. (2007) mentioned that a chatbot should deliver an effective conversation in preventing failure. Especially in natural human language, expressions diverse for various purposes. Preventing conversational errors by directing learners to the correct conversation path is critical. Therefore, fallback response should be explicitly addressed to re-engage users into the conversation (Jain et al., 2018).

2.2. The Impact of Chatbot-guided instructions on achievement

Positive effects of chatbot have been reported in research, such as longer memory retention, enhanced critical thinking skills, and improved language use and engagement (Abbasi & Kazi, 2014; Goda et al., 2014; Kerly et al., 2007; Heller, Proctor, Mah, Jewell, & Cheung, 2005; Huang, Hew, & Gonda, 2019; Wang & Petrina, 2013). Abbasi and Kazi (2014) measured the students' learning outcomes and memory retention by comparing the use of a chatbot and the Google search engine. The results showed that the students remembered the responses from the chatbot more than the Google search engine, and they outperformed the students in the Google search engine group. Goda et al. (2014) showed that English as a Foreign Language (EFL) students' critical thinking skills were also enhanced after working with the chatbot, and they were more engaged in learning. Similarly, Wang and Petrina (2013) suggested chatbot is more beneficial to intermediate or lower levels of language learners, as chatbot can be designed to repeat the same materials. These findings suggest that interaction with a chatbot serves great potential for students to engage in the learning process.

Students are motivated when they have someone to talk to during instructions. Studies indicated that a chatbot made the learning task more manageable, and the students were enjoyable by interacting with the chatbot (Heller et al., 2005; Huang, Hew, & Gonda, 2019; Kerly et al., 2007). Furthermore, Huang et al. (2019) designed three chatbots using IBM Watson Assistant (multiple-choice questions, case study, and dictionary FAQs chatbots) to assist with graduate student learning. The design of the chatbots combined video lectures, online quizzes, and answering questions. Although the majority of the students showed neutral and positive experiences with the chatbots, some had negative experiences. Those students felt the chatbot did not speak like a human. Such a result was consistent with Heller et al. (2005), and this indicated that the character of a chatbot is critical. They further noted that natural language understanding (NLU) is a significant limitation of the chatbot because open-ended questions and unstructured problems may confuse the chatbot. Consequently, Kerly et al. (2007) and Wang and Petrina (2013) advised using student-produced data (i.e., dialogues) to refine the chatbot and to overcome NLU limitation. However, their suggestion requires a large amount of student data.

Although a chatbot offers potentials for enhancing student learning, research attempting to implement a chatbot often structures the chatbot with text-based, unnatural scripts. Most importantly, the effects of a chatbot on student writing improvement are still underexplored, because many studies have focused on either language learning in general (Bii, 2013; Goda et al., 2014; Kerly et al., 2007; Wang & Petrina, 2013), or improving students' thinking process (Abbasi & Kazi, 2014; Heller et al., 2005; Huang et al., 2019).

From the technical perspective, based on the literature of the chatbot design (Jain et al., 2018), we grounded our chatbot within Jain's et al. chatbot design recommendations. Thus, we incorporated button clicking, quizzes, and question answering functions and dog pictures to help students write a thesis statement and hope this design will facilitate students' positive learning experiences.

From instructional perspectives, the chatbot was designed primarily for reinforcing process writing instructions (Flower & Hayes, 1987; Graham & Sandmel, 2011). Generating a thesis statement is a key to successful argumentative essays (De Rycker & Ponnudurai, 2011), whereas providing feedback to peers' work might improve the quality of a draft (Guardado & Shi, 2007; Rollinson, 2005). Furthermore, learning to write can be reinforced by social interaction when novice writers begin to engage the conversational process with an agent, such as a chatbot (Cazden, 1987; Kalina & Powell, 2009). Therefore, the instructional features of the chatbot used in the present study mainly reinforce a small area in the writing process, such as helping students to generate a thesis statement.

2.3. Challenges of writing thesis statements

In student writing, there is always a mismatch between their thesis statements and the supporting claims (Cekiso, Tshotsho, & Somniso, 2017; Miller & Pessoa, 2016), or sometimes their thesis statements are absent (Cekiso et al., 2017; Owusu & Yeboah, 2014). For instance, when Miller and Pessoa (2016) investigate history students' difficulties in writing thesis statements, their results indicate that the student thesis statements appear to be too general, lack contextualization, or mismatch between thesis statements and the supporting claims. Similarly, Cekiso et al. (2017) found that first-year foreign multilingual writers present similar coherence problems when they are asked to produce thesis statements on a controversial current event. The coherence problems include an absence of thesis statements, conclusions not related to the thesis statements, and confusing long sentences. Overall, the challenges of writing thesis statements seem to have a strong influence regarding its presence, rhetorical function, location, disciplinary writing practice or genre. Particularly, in writing general argumentative essays, students need to be taught that thesis statements need to be contextualized and positioned right within the introductory paragraph. Therefore, these point to a pedagogical need for a writing activity that has a strong focus on helping writers to come up with thesis statements. If a chatbot can fulfill this need, we think the chatbot will play a strong supplementary role in helping writers to develop a stronger thesis statement needed for their argumentation.

2.4. Peer review instruction

Peer review is an essential part of the writing process. Many writing instructors implement peer review as part of their writing courses. The basis of using peer review in writing instructions has two important theoretical components: process writing instructions and peer learning. In literature, peer review is part of the process writing model approach developed in the 1980s (Flower & Hayes, 1981; Keh, 1990). In the process approach, writing is a multi-staged, multi-drafted process in which students generate different versions of their work based on the feedback they receive from their peers. If necessary, the entire process could be repeated until the draft is ready as a final product.

Several studies indicate that peer review offers instructional benefits, such as reducing instructors' grading load and improving students' writing practices (Cho & Schunn, 2007; Cho & MacArthur, 2010; Cho & Cho, 2011). For instance, writing instructors adopt peer review for assessment purposes, and student writers use peer review as a reference point to improve the quality of their writing. Moreover, several studies suggest that student view peer feedback more positively than instructor feedback because peer feedback provides more extensive detailed views than instructor feedback (Cheng, Liang, & Tsai, 2015; Cho & Cho, 2011; Cho & MacArthur, 2010; Cho & Schunn, 2007; Topping, Smith, Swanson, & Elliot, 2000). Research has also found peer review activity not only enhances students' critical thinking skills but also facilitates social interactions and course engagements among peers (Kulkarni, Kotturi, Bernstein, & Klemmer, 2016). Taken together, these findings not only suggest that incorporating peer review may help to cultivate students' writing skills and reduce instructors' teaching load but also point to a need for educators to think of an innovative way to motivate writers in the peer review process.

3. Overview of the writing Chatbot DD

The learning design of the chatbot DD was based on Jain's et al. (2018) design framework. We developed the chatbot DD using Rasa (version: rasa_core 0.11.12, rasa_core_sdk: 0.11.5, rasa_nlu 0.13.7), an open-source conversational AI framework (Bocklisch, Faulkner, Pawlowski, & Nichol, 2017). We chose Rasa because Rasa emphasizes the needs of non-specialist software developers in the research field (Bocklisch et al., 2017). Figure 1

shows the design of the chatbot DD. The Rasa core module processes dialogues within a domain (the universe where the bot lives in). The Rasa core SDK module includes customizable actions (e.g., bridging the connection between Rasa core and the database). The training data for the chatbot are stored in the Rasa NLU module. The chatbot is then deployed on a web-based server using Chatroom API. The students can access the chatbot through a laptop, cellphone, or tablet, as shown in Figure 2. Following Jain's et al. (2018) suggestion when dealing with a failure conversation, the student can follow the instruction if the conversation with the chatbot encounters an error, as demonstrated in Figure 2.

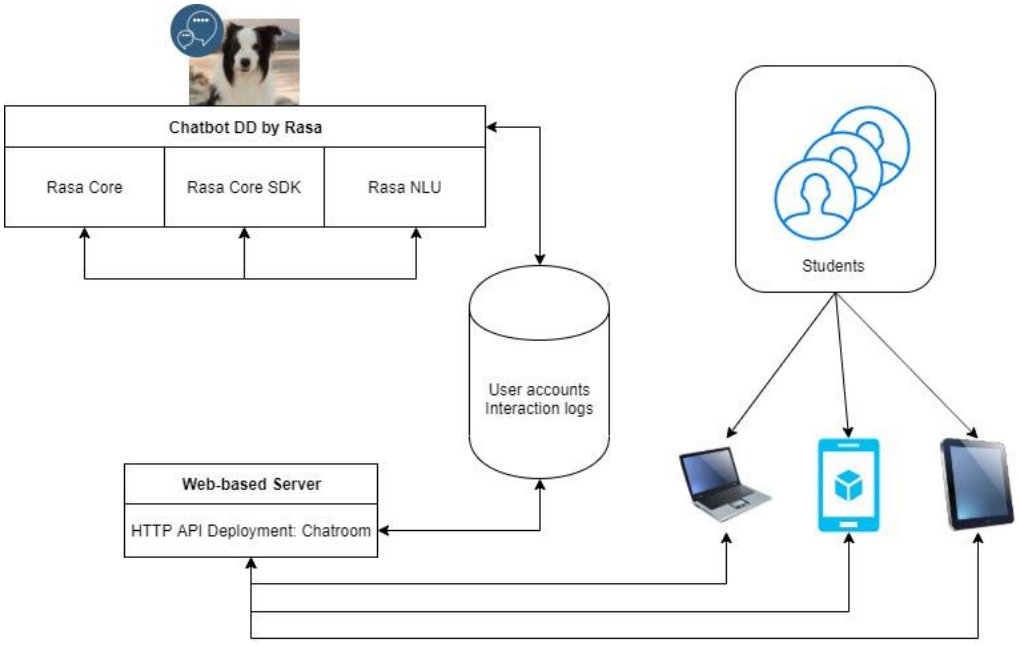


Figure 1. The framework of the chatbot DD

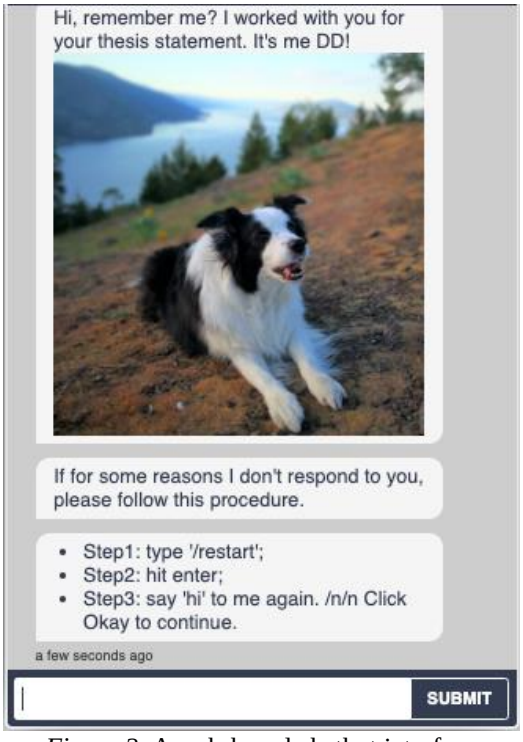


Figure 2. A web-based chatbot interface

The design of the chatbot DD was consulted with the course instructor and a former graduate-level teaching assistant. The chatbot first greets students and asks the student's identification (ID) number. This ID is stored in the database for future referencing and data analysis. Three design principles are taken from Jain's et al. recommendations (2018): field-specificity, embedded multimedia resources, and fallback response.

There were several challenges when designing the chatbot. One of the major issues is the standability of NLU. Similar phrasing, synonyms, or grammatical errors may confuse the chatbot (Clarizia et al., 2018; Huang et al., 2019; Jain et al., 2018). For example, "I want to improve my thesis statement because my teacher found I made many grammar mistakes and highlighted many errors in my essay. I don't know what to do now." Such a long sentence from a student may cause parsing errors and confusion. Because when the student says, "I want to …" and "I don't know what to do…", the chatbot may not correctly parse the student's sentence and understand whether the student needs help (I want do…) or needs clarification given pre-existing knowledge (I don't know what to do…).

Thus, to minimize the unpredictability of user input and confusion, we structured the chatbot DD by employing a button-clicking function. The fallback response will be triggered when the chatbot DD is unable to recognize a student's typed sentences. The chatbot DD recommends the student follow a certain step to re-start the program, as shown in Figure 2.

Moreover, we tried to structure the chatbot DD to speak the human-like natural language with the use of some dog pictures, which would motivate student learning and facilitate their engagement (Fryer & Carpenter, 2006; Jain et al., 2018). The scripts have been evaluated by three experienced graduate students, who have had post-secondary teaching experience, to ensure the conversation is friendly and natural. Figure 3 shows how the chatbot explained the course concept and the features of a thesis statement to the students. Figure 4 illustrates an example of student interaction. The Yes/No judgement was the concept checking questions (CCQs) after the students received lessons from the chatbot. Students can make their initial learning judgment, followed by confirmation of their judgement from the chatbot (Huang et al., 2019). All the interactions with the chatbot are automatically saved in the Rasa server.

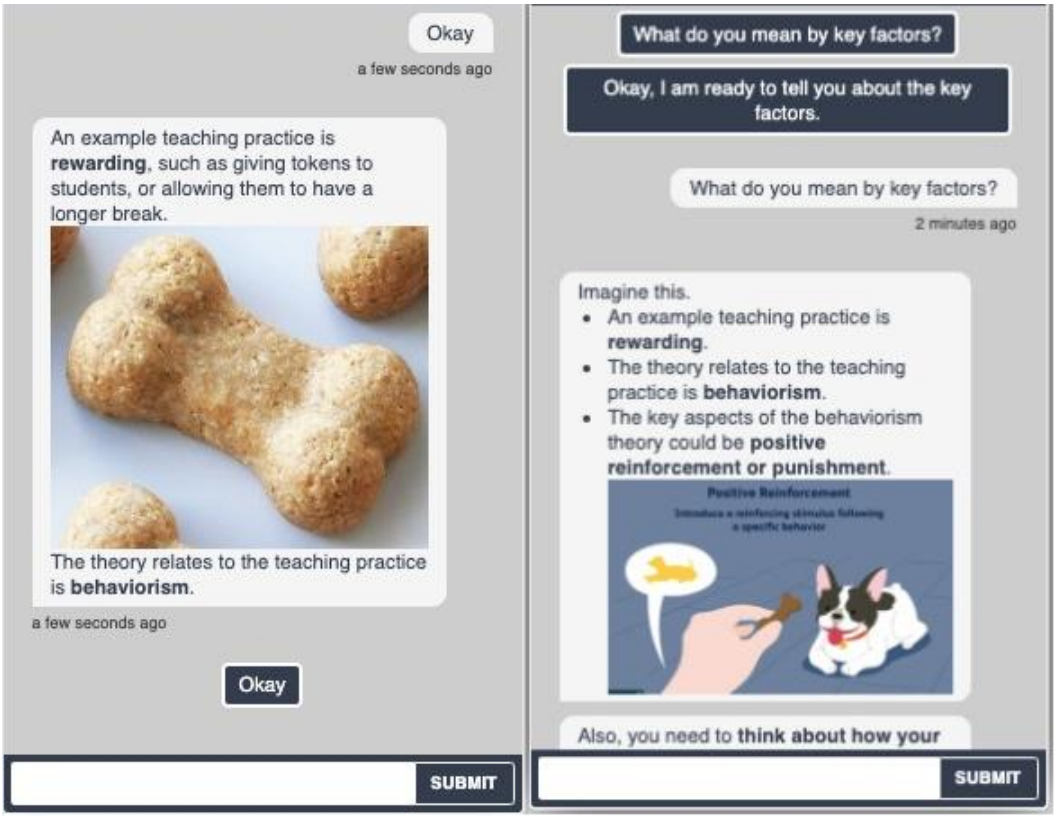


Figure 3. The chatbot DD explained the concepts of thesis statement to the students

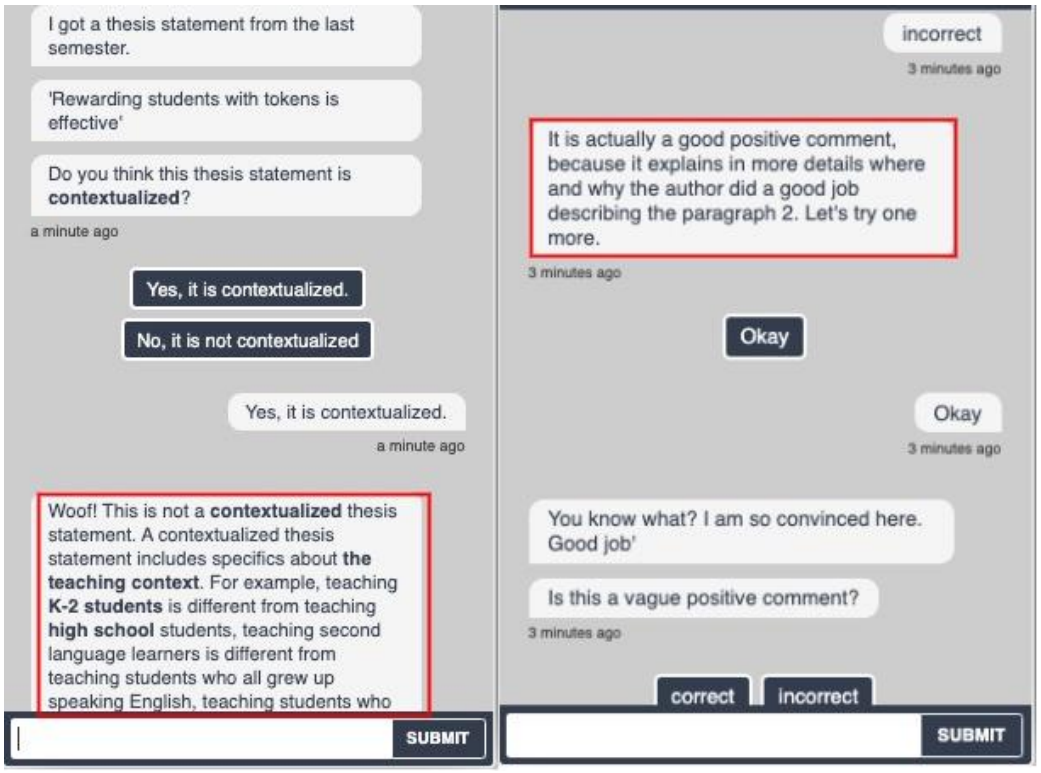


Figure 4. The chatbot tested student understanding with explanations

4. Methods

4.1. Participants

The participants were recruited from a large introductory educational psychology class (ED 100) offered in two consecutive semesters (i.e., Fall 2018 and Spring 2019) at a western Canadian university. The course did not have any prerequisites. The classes offered in two consecutive semesters were identically taught by the same instructor using the same curriculum. At this university, writing was not taught as an independent skill set, yet writing was integrated as part of a subject matter curriculum. A major flaw of this study was that no official measure of writing proficiency was administered before the intervention. The course was elective; every student could enroll if they intended to declare education as the major study subject. We ran an independent sample t-test on the midterm examination, a measure before they turned in their essay outlines. There was no statistically detectable difference between the two semesters. So, this implied that these students started at the same level of disciplinary concept knowledge before they began the essay-outline assignment.

The course had two components. One was a two-hour lecture, where the instructor taught the curriculum to the students, whereas the other was a one-hour tutorial class, where the teaching assistants led the class and answered questions about the course content and assignment expectations. Since this was a large class, there were 11 tutorials in which each tutorial had roughly 15~18 students. A teaching assistant was responsible for roughly 3~4 tutorials. There were three teaching assistants for this class in total. A semester at this university was roughly 13-week long. The students attend the two-hour lecture and one-hour tutorial each week.

The chatbot, which assisted their essay outline writing assignment, was introduced to the tutorial classes in week 6 and week 7 of Spring 2019. In week 6, the students were introduced the chatbot DD, which the chatbot DD will help their generation of thesis statements. During the week 6 tutorial, the students interacted with the chatbot in a computer lab room and came up with a thesis statement for their essay outline assignment. In week 7, the students were expected to bring a draft of their essay outline assignment to the class. The teaching assistants and the researchers led the peer review activity with the chatbot.

There were 190 undergraduate students in the Fall of 2018, and 167 students in the Spring of 2019 class. As mentioned above, in the Fall of 2018, the students were not introduced the chatbot (comparison group). In contrast, in the Spring of 2019, the chatbot was introduced to each tutorial (treatment group). With their written consent, there were 28 students (n = 28) from the spring cohort who agreed to fill out the questionnaire about their learning experiences with the chatbot DD. In order to comply with the institutional research ethics obligations, we only selected the students who have granted permission for us to analyze their questionnaire data.

4.2. Instruments

There were three instruments used in this study: the chatbot, an essay outline, and a questionnaire, respectively. First, the chatbot was designed to assist students with improving their thesis statement for the essay outline assignment. The essay outline was a graded component of the course determined by the instructor. The outline served as the planning stage for the students to draft their ideas for the final argumentative essay assignment (Flower & Hayes, 1981). In our study, the essay outline was the measure of student writing achievement. The essay outline contained several major rhetorical features of an argumentative essay, such as a thesis statement, topic sentences, evidence to support the topic sentences, counterarguments, and a conclusion. Each student needed to identify a teaching practice and argued for why the teaching practice of their choice can motivate students to learn, drawing on their course knowledge from the motivational theories of Educational Psychology. Three teaching assistants graded the essay outlines in the course. Before grading, the instructor calibrated the consistency of scoring by hosting a two-hour-long meeting. During the calibration, the teaching assistants were introduced the marking rubric of the entire outline, and then they were assigned one student essay outline for a grading attempt. The instructor repeated the process until the teaching assistants achieved consistency. The grades of the essay outline were used as the quality measure for the students in both experimental and comparison semesters. The marking rubric contained thesis statement (3 marks), arguments (3 marks), and counterarguments, including a rebuttal (4 marks). So, the essay outline was worth 10 points of the course grade. The questionnaire (see Appendix A) was adopted from Schunn, Godley, and DeMartino (2016), Topping et al. (2000), and Torrance, Thomas, and Robinson (1994). The questionnaire measured students' experience with the writing chatbot. Yes/No questions were used to avoid ambiguity in the survey statements and socially desirable responses (Mick, 1996).

4.3. Procedure

Each student in both experimental and comparison semesters was required to submit one essay outline as one of their course assignments. Each student spent two weeks working on their essay outline assignment. The students in the comparison semester wrote the essay outline without interacting with the chatbot. In the experimental semester of weeks 6 and 7, the students interacted with the chatbot to learn to construct a thesis statement for their essay outline, which was a major rhetorical feature of an essay outline. Each tutorial class was fifty minutes. During the first week of the class, drawing discipline-specific examples from Educational Psychology, the chatbot DD introduced the components of a thesis statement and guided the students to write a thesis statement for their essay outline. In the second week of the class, the students reviewed another students' outline by interacting with the chatbot DD. The chatbot DD was programmed with the ability to guide the students in providing effective peer feedback for the outline. The students then submitted their essays in week 13. At the end of the semester, the consent form was distributed to the students in the experimental semester, and they had time to fill out a questionnaire regarding their experiences with the chatbot DD until the semester ended. The questionnaire can be found in Appendix A.

4.4. Data analysis

Two types of data were collected. Qualitative data was the student responses to the open-ended questions of the questionnaire (Q33, Q36, Q37, Q38, and Q39), whereas quantitative data were (1) the student grades from the essay outline assignment and (2) the student responses to the yes/no questions of the questionnaire. Thus, data collected in this study included the score of the essay outline and a questionnaire regarding the experience with the chatbot. In this study, we used mixed methods to examine the data quantitatively and qualitatively. Studies showed that mixed methods provide a holistic and valid view by exploring the in-depth and effects of an innovative tool (i.e., a chatbot) from instructional and student perspectives in the field of educational technology (Creswell & Clark, 2017;

Randolph, 2008). An effective approach to analyze students' open-ended questions is content analysis, as this method uncovers and explores student data to generate inferences about their chatbot learning experience (Patton, 1990; Weber, 1990; Yang, 2010). For the yes/no questions from the questionnaire, we presented descriptive statistics for the student perception of the Chatbot DD. Furthermore, an independent sample t-test was used to examine the difference in writing achievement between the two semesters.

5. Results

5.1. Comparison of the writing achievements between the comparison and experimental semesters

As mentioned in the section of Participants, the students of the two groups were at the same level of disciplinary concept knowledge before they began the essay-outline assignment. Therefore, the outline scores were used to evaluate the effects of the chatbot on students writing achievements.

190 students who were in the comparison semester without using the chatbot and 167 students were in the experimental semester using the chatbot. An independent sample t-test was conducted. The outline scores in the experimental semester (M = 7.27, SD = 2.52) were statistically better than the scores in the comparison semester (M = 7.18, SD = 2.14, t = -0.38, p = 0.027*).

5.2. Student perceptions towards the Chatbot DD

Table 1 summarizes the results of students' experience with the chatbot DD. More than 80% of the students pointed out the chatbot DD helped them to identify new issues, to improve how to give effective feedback, and to become better reviewers. 78.6% of the students felt that interacting with the chatbot DD was enjoyable. 75% of the students mentioned the chatbot DD enhanced their skills in evaluating a thesis statement.

Table 1. Student experience with the writing chatbot DD

	Identified issues during peer review	Improved peer review feedback	Enjoyable	Became better reviewer	Resolve confusions of instructions	Construct precise thesis statement	Helped me evaluate my thesis statement
Yes	85.71% (24)	82.1% (23)	78.6% (22)	82.1% (23)	64.3% (18)	75.00% (21)	75.00% (21)
No	14.29% (4)	17.9% (5)	21.4% (6)	17.9% (5)	21.4% (6)	25.00% (7)	25.00% (7)
N/A	0% (0)	0.0% (0)	0.0% (0)	0.0% (0)	14.3% (4)	0.0% (0)	0.0% (0)
Total	28	28	28	28	28	28	28

Note. The number within parentheses mean the numbers of students.

Notably, when asking the students in what specific the chatbot DD helped them in writing, seventeen of the students (n = 17) indicated the chatbot DD guided them to improve thesis statement, seven of the students (n = 7) pointed out the chatbot DD enhanced their skills on giving feedback, one of the students (n = 1) felt s/he improved on both skills. Three of the students (n = 3) noted the chatbot DD did not help them at all.

The students also suggested some improvements to the chatbot DD as shown in Table 2. Nine students (n = 9) recommended the chatbot should respond faster. Seven students (n = 7) wanted the chatbot to provide more feedback and examples on the topic they learn. Four students (n = 4) noted the explanation on a thesis statement and terms definition could be simplified. Some students had negative opinions about the chatbot because sometimes the response time was slow due to networking or NLU retrieval issues. Thus, they would instead seek help from the teaching assistants or instructor.

Table 2. Student responses to the open-ended questions from the questionnaire

Themes	Findings
Favouring the Chatbot DD's overall feedback and information	*"Detail specific feedback is beneficial*" "*It was really informative"*
User-friendly interface	*"I have the flexibility to skip a lesson or choose a certain topic to learn*" "*It is easy to use*" "*I can just click the buttons instead of typing the answers*" "*It was interactive*" "*It was like chatting with a dog*" "*…talking to a dog made the activity more enjoyable and less stressful*"
Promoting learning	"*The questions DD asked and prompted made me be more critical and reflective, and it benefits in a way being self-regulated*" "*It gave me time to think and reflect on thesis statement and arguments it helped me evaluate other arguments and create an expectation for my paper*"
Positive learning experience	"*Positive experience and greatly assisted in the writing of my thesis statement and peer feedback*" "*It was a relatively positive experience. I enjoyed working with the bot and it helped me create a much better thesis statement.*" "*I think it was a fun, interactive way to improve our writing. It was something unique that I had never tried before which caught my attention!*"

6. Discussion

6.1. Do these chatbot-led writing activities enhance the students' writing achievement?

The general outcome of the present study might indicate the potential of using a chatbot as the instructional supplement to teach writing. This novel design of chatbot DD aims to supplement thesis-statement and peer-feedback instructions in our study. To our best knowledge, there has not been literature yet specifically integrating a chatbot in supporting writing instructions. However, there have been some active ITS systems that were developed to teach writing, such as the Writing Pal Intelligent Tutoring System (Roscoe, Allen, Weston, Crossley, & McNamara, 2014) and iStart (McNamara, Levinstein, & Boonthum, 2004). These ITS have been found effective and useful in teaching writing. Future research should carefully operationalize chatbot use in the context of ITS. It is because the development of an educational writing chatbot still bears some difficult technical realities, such as limitations in NLU, which students still cannot freely talk about their writing issues, and the chatbot cannot understand their problems. The NLU limitation is the reason why our chatbot has a structured dialogue. Overall, our finding needs to be interpreted with caution. First, no measure of a pre-test was administered, although the midterm examination implied that the participants were at the same level of the disciplinary concept knowledge before the treatment. Therefore, we reserve this conclusion for extended research. Second, the improved performance might be attributed to the novelty effect (Fryer, Nakao, & Thompson, 2019) – the tendency of increased achievement results from the initial introduction of an innovative technology in which users' interest level is high. In our study, the chatbot was introduced only for two weeks, so it was likely that the students felt very curious, and they showed a strong interest in using the chatbot. Future research might consider observing whether performance still sustained if the chatbot was introduced for a longer period.

6.2. How do students perceive the use of chatbot in a university classroom

When interacting with the chatbot, among those who responded to the questionnaire, 75% of them felt that DD helped them write a precise thesis statement and evaluate the quality of the thesis statement. These findings add

further support to the supplementary role of technology in writing instructions and the importance of conversation when writers are learning to write (Bii, 2013; Kalina & Powell, 2009). Furthermore, consistent with previous research reporting positive student experiences with a chatbot (Fryer & Carpenter, 2006; Goda et al., 2014; Kerly et al., 2007; Luger & Sellen, 2016), we have found that approximately 79% of the students reported an enjoyable experience with the chatbot. Our finding might suggest that integrating a chatbot with writing instruction might improve student learning-to-write engagement.

Asking students to review their peers' essays requires extensive instructional explanations (Cho & Schunn, 2007; Li, Liu, & Steckelberg, 2010; Yang, 2011). In our study, 85% of the students reported the chatbot helped them identify their writing issues; 82% of the students felt they become a better reviewer with improved peer feedback quality, and 64% of the students mentioned the chatbot helped them to resolve confusions of the peer review instructions. Therefore, using a chatbot on the side of a peer review activity might solve instructional confusion and benefit students in becoming better peer reviewers.

On the design and development side of the chatbot, it seems the students enjoyed learning with the chatbot because the chatbot has multimedia features, such as using an animal dog picture, friendly language, and button clicking. From the open-ended survey data, some students liked the button clicking feature as it made the chatbot user-friendly. Furthermore, some students liked the chatbot was embodied by a dog picture, which made the overall learning experience more entertaining. These findings are consistent with previous research in educational chatbot design that directive, friendly language and the button clicking function are crucial in engaging student learning and minimize the difficulties of learning tasks (Heller et al., 2005; Huang et al., 2019; Jain et al., 2018; Kerly et al., 2007).

7. Conclusion and limitations

In conclusion, a writing chatbot was introduced as part of a disciplinary writing class. Although some positive student testimonials and improved essay outline performance were reported, the present study has some limitations. Particularly, the writing proficiency of the participating students was not well-controlled. In the present study, administering a pre-test was not administered. Future research examining the effect of a chatbot on writing achievement need to take students' writing proficiency into account for analysis. Secondly, the improved performance of the essay outline might be due to the novelty effect. Future researchers might extend the use of chatbot over a longer period to see if performance and student interest levels still sustain (Fryer et al., 2019). Despite several methodological concerns, we argue that chatbot has by far used as a supplement instead of a standalone ITS because of its NLU constraint (Heller et al., 2005).

Also, we developed the chatbot based on the existing design framework, including external database storage, conversation failure mechanism, images and quizzes combination, delivering effective and natural conversation, and user-friendly interface. What advances the current research will be collecting and establishing a large-scale repository of student writing issues and its solutions to develop a classification system that correctly identifies a student's learning dialogue (Huang et al., 2019; Jain et al., 2018; Wang & Petrina, 2013). Future chatbot designers and researchers are recommended to overcome the limitation of NLU, so students can freely talk to a chatbot without encountering conversational errors, or fallback.

Acknowledgement

This research was funded by Dr. Phil Winne's research grant in support of ML's doctoral research assistantship (Social Sciences and Humanities Research Council SSHRC#435-2015-0273). Parts of this research were partitioned into the authors' doctoral dissertations. We thank all the research assistants in Dr. Winne's lab for the conceptual development of the chatbot. We also thank Mr. Kenny Teng for providing technical assistance during the class sessions. We also thank the Editor of the journal and the two anonymous reviewers for offering the feedback. All names/courses/institution names have been mocked to ensure confidentiality.

References

Abbasi, S., & Kazi, H. (2014). Measuring effectiveness of learning chatbot systems on student's learning outcome and memory retention. *Asian Journal of Applied Science and Engineering*, *3*(2), 251-260. doi:10.15590/ajase/2014/v3i7/53576

Bii, P. (2013). Chatbot technology: A Possible means of unlocking student potential to learn how to learn. *Educational Research*, *4*(2), 218-221.

Bocklisch, T., Faulkner, J., Pawlowski, N., & Nichol, A. (2017). *Rasa: Open source language understanding and dialogue management*. arXiv:1712.05181 [cs.CL]

Cazden, C. B. (1988). *Classroom discourse: The language of teaching and learning*. Portsmouth, NH: Heinemann.

Cekiso, M., Tshotsho, B., & Somniso, M. (2017). Exploring first-year university students' challenges with coherence writing strategies in essay writing in a South African university. *International Journal of Educational Sciences*, *12*(3), 241-246. doi:10.1080/09751122.2016.11890431

Cheng, K. H., Liang, J. C., & Tsai, C. C. (2015). Examining the role of feedback messages in undergraduate students' writing performance during an online peer assessment activity. *The Internet and Higher Education*, *25*, 78-84. doi:10.1016/j.iheduc.2015.02.001

Cho, Y. H., & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science*, *39*(5), 629-643. doi:10.1007/s11251-010-9146-1

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction*, *20*(4), 328-338. doi:10.1016/j.learninstruc.2009.08.006

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A Web-based reciprocal peer review system. *Computers & Education*, *48*(3), 409-426. doi:10.1016/j.compedu.2005.02.004

Clarizia, F., Colace, F., Lombardi, M., Pascale, F., & Santaniello, D. (2018). Chatbot: An Education support system for student. In *International Symposium on Cyberspace Safety and Security* (pp. 291-302). Amalfi, Italy: Springer. doi:10.1007/978-3-030-01689-0_23

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage publications.

De Rycker, A., & Ponnudurai, P. (2011). The Effect of online reading on argumentative essay writing quality. *GEMA Online® Journal of Language Studies*, *11*(3), 147-162.

Edwards, D., & Mercer, N. (2013). *Common Knowledge (Routledge Revivals): The Development of understanding in the classroom*. London, United Kingdom: Routledge.

Flower, L., & Hayes, J. R. (1981). A Cognitive process theory of writing. *College composition and communication*, *32*(4), 365-387. doi:10.2307/356600

Fryer, L., & Carpenter, R. (2006). Bots as language learning tools. Language learning and technology. *Language Learning & Technology*, *10*(3), 8-14. doi:10125/44068

Fryer, L. K., Nakao, K., & Thompson, A. (2019). Chatbot learning partners: Connecting learning experiences, interest and competence. *Computers in Human Behavior*, *93*, 279-289. doi:10.1016/j.chb.2018.12.023

Goda, Y., Yamada, M., Matsukawa, H., Hata, K., & Yasunami, S. (2014). Conversation with a chatbot before an online EFL group discussion and the effects on critical thinking. *The Journal of Information and Systems in Education*, *13*(1), 1-7. doi:10.12937/ejsise.13.1

Ghose, S., & Barua, J. J. (2013). Toward the implementation of a topic specific dialogue based natural language chatbot as an undergraduate advisor. In *2013 International Conference on Informatics, Electronics and Vision (ICIEV)* (pp. 1-5). Dhaka, Bangladesh: IEEE. doi:10.1109/ICIEV.2013.6572650

Graham, S., & Sandmel, K. (2011). The Process writing approach: A Meta-analysis. *The Journal of Educational Research*, *104*(6), 396-407 doi:10.1080/00220671.2010.488703

Guardado, M., & Shi, L. (2007). ESL students' experiences of online peer feedback. *Computers and Composition*, *24*(4), 443-461. doi:10.1016/j.compcom.2007.03.002

Heller, B., Proctor, M., Mah, D., Jewell, L., & Cheung, B. (2005, June). Freudbot: An Investigation of chatbot technology in distance education. In *EdMedia+ Innovate Learning* (pp. 3913-3918). Waynesville, NC: Association for the Advancement of Computing in Education (AACE).

Huang, W., Hew, K. F., & Gonda, D. E. (2019). Designing and evaluating three chatbot-enhanced activities for a flipped graduate course. *International Journal of Mechanical Engineering and Robotics Research*, *8*(5). doi:10.18178/ijmerr

Jain, M., Kumar, P., Kota, R., & Patel, S. N. (2018). Evaluating and informing the design of chatbots. In *Proceedings of the 2018 on Designing Interactive Systems Conference 2018* (pp. 895-906). Hong Kong, China: ACM. doi:10.1145/3196709.3196735

Kalina, C., & Powell, K. C. (2009). Cognitive and social constructivism: Developing tools for an effective classroom. *Education*, *130*(2), 241-250.

Keh, C. L. (1990). Feedback in the writing process: A Model and methods for implementation. *ELT Journal*, *44*(4), 294–304. doi: 10.1093/elt/44.4.294

Kerly, A., Hall, P., & Bull, S. (2007). Bringing chatbots into education: Towards natural language negotiation of open learner models. *Knowledge-Based Systems*, *20*(2), 177-185. doi:10.1016/j.knosys.2006.11.014

Kirakowski, J., O'Donnell, P., & Yiu, A. (2009). Establishing the hallmarks of a convincing chatbot-human dialogue. *Human-Computer Interaction*, 49-56. doi:10.5772/7741

Kulkarni, C., Kotturi, Y., Bernstein, M. S., & Klemmer, S. (2016). Designing scalable and sustainable peer interactions online. In *Design Thinking Research* (pp. 237-273). doi:10.1007/978-3-319-40382-3_14

Li, L., Liu, X., & Steckelberg, A. L. (2010). Assessor or assessee: How student learning improves by giving and receiving peer feedback. *British journal of educational technology*, *41*(3), 525-536. doi:10.1111/j.1467-8535.2009.00968.x

Luger, E., & Sellen, A. (2016). Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5286-5297). San Jose, CA: ACM. doi:10.1145/2858036.2858288

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), 901-908. doi:10.1037/a0037123

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 222-233. doi:10.3758/BF03195567

Mick, D. G. (1996). Are studies of dark side variables confounded by socially desirable responding? The Case of materialism. *Journal of consumer research*, *23*(2), 106-119. doi:10.1086/209470

Miller, R. T., & Pessoa, S. (2016). Where's your thesis statement and what happened to your topic sentences? Identifying organizational challenges in undergraduate student argumentative writing. *TESOL Journal*, *7*(4), 847-873. doi:10.1002/tesj.248

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

Owusu, E., & Adade-Yeboah, A. (2014). Thesis statement: A Vital element in expository essays. *Journal of Language Teaching & Research*, *5*(1), 56-62.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.

Pereira, J., & Díaz, Ó. (2018). Chatbot dimensions that matter: Lessons from the trenches. In *International Conference on Web Engineering* (pp. 129-135). Cáceres, Spain: Springer. doi:10.1007/978-3-319-91662-0_9

Randolph, J. J. (2008). *Multidisciplinary methods in educational technology research and development*. Hämeenlinna, Finland: HAMK Press/Justus Randolph.

Rodríguez-Gil, L., García-Zubia, J., Orduña, P., Villar-Martinez, A., & López-De-Ipiña, D. (2019). New approach for conversational agent definition by non-programmers: A Visual domain-specific language. *IEEE Access*, *7*, 5262-5276. doi:10.1109/ACCESS.2018.2883500

Rollinson, P. (2005). Using peer feedback in the ESL writing class. *ELT journal*, *59*(1), 23-30. doi:10.1093/elt/cci003

Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2014). The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition*, *34*, 39-59. doi:10.1016/j.compcom.2014.09.002

Schunn, C., Godley, A., & DeMartino, S. (2016). The Reliability and validity of peer review of writing in high school AP English classes. *Journal of Adolescent & Adult Literacy*, *60*(1), 13-23. doi:10.1002/jaal.525

Song, D., Oh, E. Y., & Rice, M. (2017). Interacting with a conversational agent system for educational purposes in online courses. In *2017 10th international conference on human system interactions (HSI)* (pp. 78-82). Ulsan, South Korea: IEEE. doi:10.1109/HSI.2017.8005002

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment & evaluation in higher education*, *25*(2), 149-169. doi:10.1080/713611428

Torrance, M., Thomas, G. V., & Robinson, E. J. (1994). The Writing strategies of graduate research students in the social sciences. *Higher education*, *27*(3), 379-392. doi:10.1007/BF03179901

Vanlehn, K. (2006). The Behavior of tutoring systems. *International journal of artificial intelligence in education*, *16*(3), 227-265.

Wang, Y. F., & Petrina, S. (2013). Using learning analytics to understand the design of an intelligent language tutor–Chatbot Lucy. *International Journal of Advanced Computer Science and Applications*, *4*(11), 124-134. doi:10.14569/IJACSA.2013.041117

Weber, R. P. (1990). *Basic content analysis* (2nd ed.). Newbury Park, CA: Sage.

Weizenbaum, J. (1966). ELIZA-a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45. doi:10.1145/365153.365168

Xu, D., & Wang, H. (2006). Intelligent agent supported personalization for virtual learning environments. *Decision Support Systems*, *42*(2), 825-843. doi:10.1016/j.dss.2005.05.033

Yang, Y. F. (2010). Students' reflection on online self-correction and peer review to improve writing. *Computers & Education*, *55*(3), 1202-1210. doi:10.1016/j.compedu.2010.05.017

Yang, Y. F. (2011). A Reciprocal peer review system to support college students' writing. *British Journal of Educational Technology*, *42*(4), 687-700. doi:10.1111/j.1467-8535.2010.01059.x

Appendix A. Questionnaire regarding student experience with the chatbot

1. Student ID: _____
2. Name: _____
3. Email: _____
4. Major: _____
5. Academic residency: _____
6. EAL: _____
7. In the past 6 months, which of these forms of writing have you engaged in? (Please select all that apply):
   a. Plans and notes (taking notes in class)
   b. Reports and assigned work (writing emails, cover letters for jobs)
   c. Writing for publication (e.g., writing a book or blog posts)
   d. Research term paper within a course (e.g., Literature review for PHIL 120; persuasive paper for FALx99)
   e. High school essays or provincial exams (e.g., English 12 or equivalent)
   f. Other (please specify): _____
8. When you write, what strategies do you always adopt or use. (choose up to 3):
   a. Brainstorming
   b. Taking notes from research sources
   c. Mindmapping
   d. Ordering notes
   e. Making an outline
   f. Drafting
   g. Revising
   h. Sharing ideas with a friend and receiving feedback
   i. Other
9. When you revise your paper, what are your goals? (choose up to 3)
   a. Improving clarity
   b. Improving style
   c. Developing content
   d. Correcting errors
   e. Rearranging the text
   f. Reducing length

10. Generally speaking, at what point do you like to start writing?
11. Structuring my arguments for my term paper is relatively easy for me (Y/N)
12. I would describe myself as a poor writer (Y/N)
13. I find writing a thesis statement is difficult (Y/N)
14. I worry a lot about whether the grammar and spelling is correct for my thesis statement (Y/N)
15. I worry that my difficulty with writing will jeopardize completing my essay (Y/N)
16. I am not good at coming up with a thesis statement (Y/N)
17. I gain a great deal of pleasure from writing (Y/N)
18. I find writing a frustrating process (Y/N)
19. I worry the clarity of my thesis statement will affect my paper grade (Y/N)
20. I find the process of writing highly stressful (Y/N)
21. I find writing a thesis statement pretty frustrating (Y/N)
22. I find the process of coming up with a thesis statement quite stressful (Y/N)
23. The easiest part of the writing process is producing a plan (Y/N)
24. Structuring my arguments to form a well-structured thesis statement is relatively easy for me (Y/N)
25. I find writing hard work (Y/N)
26. I worry a lot about whether my grammar and spelling are correct (Y/N)
27. Working with DD the Thesis Bot helped me construct my thesis statement more precisely (Y/N)
28. Working with DD helped me evaluate my own thesis statement (Y/N)
29. The bot DD helps me identify new issues with my peer's dialectical map (Y/N)
30. The bot DD helps improve my feedback for my peer's dialectical map (Y/N)
31. What specifically did DD help me? (Open-ended)
32. When I worked with DD, he helped me change the quality of my review (Y/N)
33. Tell me how did (or did not) your review change because you engaged with the bot (Open-ended)
34. Working with DD the Peer Review/Thesis Bot was enjoyable (Y/N)
35. Working with DD taught me how to be a better reviewer of my peer's work (Y/N)
36. Did you find working with DD the Peer Review/Thesis Bot helped you resolve some confusions from the instructions? If yes, what specifically was resolved? (Open-ended)
37. Describe the experience you have had in the peer-review/thesis chatbot (Open-ended)
38. What did you like about the DD chatbot? (Open-ended)
39. What suggestions would you provide to make the chatbot more effective? (Open-ended)