Testing for a Test Mode Effect: A Quasi-Experimental Study Using EFL Vocabulary Quizzes

Ross Eric Miller

Abstract

The present study used a quasi-experimental design to look for a potential test mode effect on English vocabulary quizzes for a required English communication course in a Japanese university. In this study, two sections of the same course took traditional paper-based vocabulary quizzes at the start of every class while one section took the same quizzes on their smartphones through a course shell made on the LMS Canvas. Scores were collected and compared from two regular vocabulary quizzes (n=93, n=102). In addition, a review quiz (n=102) was also administered in order to check for a possible effect the testing mode would have on recall of previously tested material. One-way ANOVA tests found no statistically significant differences between any of the groups on any of the measures. The discussion highlights the need for practitioners to consider the potential for unintended consequences that can occur when technology is unevenly integrated among students within the same program.

Introduction

Technology in teaching is a two-way street that provides students with different opportunities for learning on the one side as it transforms the teaching experience for educators on the other (Aldunate & Nussbaum, 2013). Technology is often seen as a means to facilitate student-centered learning (Davies & West, 2014; McKnight et al., 2016). However, in practice, it has been found that even in teachers whose pedagogical beliefs align with constructivist practices, implementation has often perpetuated traditional teacher-centered practices (Ertmer et al., 2012; Ertmer & Ottenbreit-Leftwich, 2010) or been simply used to enhance instruction (Spires & Bartlett, 2012).

But not all technology is employed for the purpose of enhancing instruction. What about the situation where technology has been adopted for practical reasons rather than pedagogical ones? One theme that has emerged as a practical reason for technology adoption is that of teachers saving time (Liu, 2011; Spotts, 1999). This is understandable because for teachers, time is often a luxury that is in short supply. While

very few professional educators would argue the use of technology for technology's sake is a good thing, many would most likely feel that any technology that makes clerical aspects of teaching easier can't be bad (Straub, 2009). However, is it possible that technology integration for the benefit of the teacher has some unintended consequences on the students? To be more specific, could a technology that is adopted to facilitate the distribution and grading of quizzes for a teacher of one group of students result in scores that are significantly different from those whose teachers are using paper-based versions of the same quiz? The necessity to investigate this question lies in the desire to ensure equality within a program where students enrolled in different sections of the same course are evaluated based on the same measures.

Literature Review

Cognitive Load

Cognitive Load refers to the mental energy required of the learner to complete a specific task. The total cognitive load has been sub-divided into three varieties, with intrinsic load, germane load and extraneous load all contributing to overall task difficulty (Antonenko & Niederhauser, 2010; Richey & Klein, 2014; Sweller et al., 1998). Intrinsic load refers to the attributes required for task completion. Germane load refers to characteristics the learners bring to the task from their individual experiences and the resultant cognitive effort required to complete the task. Extraneous load is the cognitive processing power that is required to complete the task, but not an inherent quality of the task itself (Antonenko & Niederhauser, 2010; Sweller et al., 1998). When a task has been designed to limit extraneous cognitive load while putting demands on germane cognitive load, better learning outcomes can be expected (Sweller et al., 1998). Test difficulty is a result of the interaction between the questions and what the learners know (intrinsic load and germane load). Teachers should endeavor to make sure that "the mode of testing does not add to the cognitive workload of the individual" (Noyes et al., 2004, p. 112). Making a test fair for all students who take it means that any potential unrelated variables that add on to extraneous cognitive load for a group of students need to be mitigated. Due to its ubiquity in modern education, one possible source of extraneous cognitive load that has the potential to be overlooked in course design is the use of online testing, which brings with it different challenges for test-takers than traditional paper-based tests.

Media in Education

The media methods debate has been focused on the difference in learning outcomes that can potentially arise from using different tools for instruction. In 1983, Clark began the debate by declaring that "the best current evidence is that media are mere vehicles that deliver instruction but do not influence student

achievement any more than the truck that delivers our groceries causes changes in our nutrition" (p. 445). Kozma (1991) countered that view, arguing that media does have an impact on the construction of knowledge because learning does not occur as a result of the delivery of instruction, but by the students' interaction and collaboration with instruction and the method in which it was delivered. When changing the perspective from one of media used to deliver content to one of media used to evaluate student performance: provided the activities are the same and the cognitive load is the same, it should be expected that the results will be the same as well (Clark, 1994).

In looking at the effect of media on learning, Hastings & Tracey (2004) add modern computers to the debate and claim that the features of the computer, especially once connected to the store of knowledge contained on the internet do make for a different kind of media that supports Kozma's view. In a 2006 article, Oblinger & Hawkins make an argument that media does make a difference in direct relation to its impact on the aspects of learning, which include such components as motivation and the ability to interact with others. However, they declare that the ability to discern any significant impact of technology is in direct relation to the clarity of the question asked: With this study, the focus is related technology implementation for evaluation purposes as opposed to knowledge acquisition, and the question we are asking is whether the testing mode (paper-based vs. online quizzes accessed with a smartphone) has the potential to impact scores on quizzes of foreign language vocabulary.

Paper-based vs. Digital Assessments

For tests items with limited answer options, the automation of an online format makes the job of grading much easier for teachers (Čandrlić et al., 2014). Does the shift to an online format for testing also make it easier for students, or is there a potential to make it more difficult? The literature provides no clear answer, perhaps because the variables involved in any test differ from teacher to subject to class to learner making it difficult to generalize.

One comparison study of online versus paper-based testing found no significant difference in scores on a 10-item comprehension test given after reading a short passage (Noyes et al., 2004). However, the researchers did find that there was statistically significant difference between the two groups regarding the perception of effort required to complete the test. Noyes et al. (2004) concluded that computer-based tests required more effort, and as a result, students who find the test questions to be more challenging will be disadvantaged if their test is computer-based.

In a comparison study of 105 students in university Computer Fundamentals course, Clariana & Wallace (2002) used 100 item multiple choice test. They highlighted the differences between their two testing formats, describing the paper-based version as having up to seven questions on a page and a separate sheet

for students to mark their answer choices from A to D. The online version had one question per screen which advanced when an answer option was selected, though students could go back to review or change answers. Upon comparing the results of the two groups, they found that the experimental group using computers had statistically significant higher mean scores. However, through further analysis using different surveys, they ruled out "computer familiarity, gender, and competitiveness" (p. 598) as explanations, and concluded that the variable most responsible for the difference in scores was content familiarity.

Čandrlić et al. (2014) conducted a study between online and paper-based tests, initiated in part to assess the viability of online testing for providing an easier way for teachers to grade. They compared the results of more than 1,200 paper-based and online tests of students from three different majors within the University of Rijeka's Department of Informatics. For this study, the tests were not always identical instruments, with paper-based tests having more essay-type questions. There were no statistical differences within two of the majors, but in the third, one group did outperform their paper-based counterparts. In that case, the online tests included only items which could be objectively graded, e.g., true/false and multiple-choice questions while the paper-based version included short answers and essays. The difference in scores was attributed to the difference in test instruments, in which the online version was deemed easier because the items could "be answered even with lower levels of knowledge and based on recognition" (Čandrlić et al., 2014, p. 661). As a result, they recommended that in order for a test to allow for a true demonstration of learning, online tests should include about 30% of essay-type questions that would need to be graded subjectively, arguing that "this ratio does not represent great time workload for the teacher" (p. 662). They also found that their students preferred keyboards over pencils when testing.

A study conducted with 251 students in a university Spanish class also found no statistically significant differences between those who took quizzes online and those who took the paper-based quizzes (Vanpatten et al., 2015). A total of three quizzes were analyzed using two methods: one with test mode being the independent variable and the scores being the dependent variable and another which used the sections as the independent variable and the three quizzes combined mean scores as the dependent variable. For this study, the researchers stressed that "the quizzes taken online and the quizzes taken in class were identical" (p. 663), so any differences would be the result of the mode used for testing. They also note that the quizzes were composed of primarily multiple choice and true/false questions which allowed them "to be easily graded by instructors" (p. 663).

Testing Effect and the Testing Media

While testing is most often used as an instrument for students to demonstrate their learning at a specific point in time, tests have also been shown to help learners retain material for future recall due to the so-

called testing effect (McDaniel et al., 2007). In research on input modes, Mangen, Anda, Oxborough, & Brønnick (2015) found that when it came to the remembrance of vocabulary items, "there may be certain cognitive benefits to handwriting which may not be fully retained in keyboard writing (p. 239)," and explained these potential benefits as the result of the cognitive/muscular control differences between handwriting and using a keyboard. A decade earlier, Longcamp, Boucard, Gilhodes, & Velay (2006) had the same results and explained the potential difference as due to handwriting requiring the visualization and recreation of the shape of a character where using a keyboard simply requires knowing the location of the proper key to push.

Theoretical Framework

The theoretical framework guiding the first part of this research is the test mode effect. The test mode effect refers to the differences in results found between identical tests when one is paper-based and the other is computer-based (Clariana & Wallace, 2002; Noyes et al., 2004). When it comes to testing, the extraneous load on students is potentially different when one assessment is paper-based and the other is digital. With digital tests, some variance in testing scores could be due to the fact that some students will be more familiar with computers than others, a condition that students are aware of and perceive of as unfair (Dashtestani, 2015).

Aside from familiarity with computers, the presentation of these two modes of testing is different. Considering identical tests between digital and paper-based media, Clariana and Wallace (2002) felt that potential differences in fonts, font sizes, screen brightness and differences in resolution might be factors in any test-mode effect. They also mentioned that differences of dimensions between a sheet of paper and the display of the computer screen could affect student perceptions of test difficulty. On mobile devices, the relatively small size of the screen and of the keyboard have been cited as limitations when using mobile technology (Dukic et al., 2015; Gitsaki & Robby, 2014). However, those qualities refer to the presentation of the test and not the actions required to complete it.

A dissection of the tasks required to complete paper-based/digital assessments illustrates the differences of extraneous load between these two modes of testing. On a paper-based assessment, assuming standard font sizes and quality of print, students must simply start at the top of the page, read the items and answer them with pens or pencils while they move down the page until finished. As the method of taking this paper-based test has been automated through years of formal schooling, the only difficulty or potential stressor with such an assessment is the difficulty of the questions (the intrinsic cognitive load), a broken pencil tip, or a lack of time. With a digital assessment, assuming the use of an LMS on a smartphone,

students must have a charged device, make sure they have a stable internet connection, then open a browser window, access the correct URL, log in with the correct ID and password, access the assessment page, click to start, decide on the proper display adjustments for optimal viewing, scroll through the page, click into the answer field, input text with a digital keyboard, click out of the answer field, and repeat scrolling, clicking, using a digital keyboard, and finally submitting when done, all before the quiz times out. The potential for losing time because of a hiccup in the process they must complete before even accessing the quiz can become an added stressor, and the possibility of not seeing a question is also higher on a digital assessment because if all of the question items are on one digital page, it becomes much easier to scroll over an item on a smartphone screen than it would be to miss one on a sheet of paper. Furthermore, a student never has to restart a test because the page of paper suddenly "crashes," something which can happen when the assessment is online.

This study looked at two potential aspects of the test mode effect. The first part examines if scores on vocabulary quizzes are impacted by the media in which they are administered. The second part of this research examines if testing modes affect future recall. Regarding the test mode effect, the range of results shown in the literature highlight the need to look for a potential effect when technology is integrated into a specific context.

Research Questions:

The focus of this research is to find out if there were significant differences between one class of students who were given digital quizzes which they took using their mobile devices and other classes of students who took the same quizzes using analog technology. Specifically, the research questions posed in this study were:

1) Does student use of mobile devices when taking vocabulary quizzes account for an advantage/ disadvantage over students who are taking paper-based versions of the same quizzes?
2) Does student use of mobile devices for taking vocabulary quizzes negatively impact their retention of previously tested items?

Methodology

Participants and context

The participants were 102 second-year students enrolled in three sections of a required English communication course at a private Japanese university in the spring of 2018. Students were placed in each

section based on the results of Test of English for International Communication (TOEIC®) given at the end of their first year of university study (January 2018). Each section was composed of roughly 25% of students from each of the four quartiles of TOEIC® results for that cohort of students. Expectations were that each section of the course would then be equally mixed in terms of the overall English abilities of the students, resulting in no discernable difference in overall ability between course sections. This was in contrast to previous years where sections were based solely on TOEIC® scores, making each section noticeably different from another in terms of English/academic ability at the beginning of the term.

These face-to-face classes met for 90 minutes twice a week over a 15-week semester. Ten percent of students' final grades were based on vocabulary quizzes. While this vocabulary was not specifically/explicitly taught during class, students were provided with a vocabulary book composed of common words/phrases that appear in the TOEIC®. Each double page spread of this vocabulary book was made up of 10 new vocabulary words appearing in short phrases along with their inflected endings, Japanese translations of the phrase/sentence, pronunciation of the item in a phonetic alphabet and an explanation of the word and its usage in Japanese. Students were given a schedule and expected to learn 40 vocabulary words a week: 20 by the first class (Mondays) and another 20 by the second class of the week (Thursdays). The quiz for each class meeting used a random sampling of 10 of the 20 new vocabulary words students were supposed to have learned for that day. The first 10 minutes of each class were set aside for these quizzes.

Materials

The format of the usual vocabulary quizzes exactly matched the formatting and presentation from the vocabulary book: the phrase with the first letter of the target word followed by a blank. The Japanese translation was also included. Inflected endings for the words, if included in the example used in the book, were also included in the quizzes. Each quiz had 10 items and was worth 10 points. Answers were worth 1 point with no partial credit given.

Items used on the review quiz were taken from the 50 words that had been used in the first five vocabulary quizzes and three question types were used. Section 1 (10 items) was a random selection of 2 words from each of the first five quizzes. No changes were made to these items. Section 2 was composed of one word (5 items) from each the first five quizzes, but in this section, the Japanese translation was removed. The third and final section of the review quiz had 6 fill-in-the-blank sentences (6 items) using vocabulary from the first five quizzes. Each sentence was crafted to provide meaningful context for the target word. So, while sections one and two could be answered from memory, section three required a display of deeper understanding of the meaning. No Japanese translation was provided for this section. This quiz had a maximum possible score of 21 points, with no partial points given. The grade for this review

quiz was not counted as part of students' grades for this course and was intended to see if test mode had an effect on retention of previously tested items.

For the regular quizzes and all of the sections of the review quiz, the first letter of the intended vocabulary word was provided. Quiz 6 and the Review Quiz are shown in the Appendices.

Procedure

This was a quasi-experimental design because the experimental group and control groups were not assigned randomly (Creswell, 2014). Three sections of this required English communication course were given vocabulary quizzes at the start of every class (twice a week). The two control groups (Paper1 and Paper2) took their quizzes on printed handouts. The experimental group (Canvas1) took their quizzes with their personal smartphones. For the experimental group, the quizzes were hosted in a course shell within the LMS Canvas that the instructor had set up with a private account. Students were not using the mobile Canvas app, so had to log in through a web browser by accessing the proper URL. Quizzes were distributed/accessible at the start of class, and students had to finish within the first 10 minutes of class time.

In this study, scores from Quiz 6 (3rd week of semester) and Quiz 22 (13th week of the semester) were collected. The review quiz was given at the start of week four. All students took a paper-based version of the review quiz to eliminate any differences due to test mode. Once collected, the scores were analyzed using the JASP statistical software package (version 0.14.1) for Apple computers. In addition to descriptive statistics, one-way ANOVA tests were conducted to see if there were any significant statistical differences between the groups. For the review quiz, analyses were conducted for each quiz section in addition to the final score.

Results

Quiz 6 (n=93)

Results from vocabulary Quiz 6 can be seen in Table 1. Scores were compared among three sections of the Communication class (n=93). The quiz was made up of items taken from the vocabulary sequenced 101 to 120 in the associated vocabulary book. For this 10-item quiz, there were a total of 10 points possible. Of the three groups, the smartphone group (n=29) had the lowest mean score of 8.52. The other two sections which completed the paper-based quiz had means of 9.30 (n=30) and 9.06 (n=34) respectively. A one-way ANOVA found no significant differences between the 3 sections: $F(2, 90) = 1.88$, $p = .159$.

Table 1 Descriptive statistics for Quiz 6 Results (n=93)

Group	Min. Score	Max. Score	Mean	SD
Paper_1 (n=30)	6.00	10.00	9.30	1.09
Paper_2 (n=34)	5.00	10.00	9.06	1.35
Smartphone (n=29)	1.00	10.00	8.52	2.18

Quiz 22 (n=102)

Results from vocabulary Quiz 22 can be seen in Table 2. Scores were compared among three sections of the Communication class (n=102). The quiz was made up of items taken from the vocabulary sequenced 421 to 440 in the associated vocabulary book. This quiz was also worth 10 points. As with Quiz 6, the smartphone group (n=32) had the lowest mean score (7.47). The other two groups which completed the paper-based quiz had means of 7.75 (n=36) and 7.59 (n=34) respectively. As with the earlier quiz, there were no statistically significant differences between the three groups. The results of a one-way ANOVA were $F(2, 99) = 0.84$, $p = .920$.

Table 2 Descriptive statistics for Quiz 22 Results (n=102)

Group	Min. Score	Max. Score	Mean	Std. Dev
Paper_1 (n=36)	1.00	10.00	7.75	2.25
Paper_2 (n=34)	1.00	10.00	7.59	2.94
Smartphone (n=32)	0.00	10.00	7.47	3.32

Review Quiz

Section 1 of the review quiz was worth 10 points, Section 2 was worth five points, and Section 3 was worth six points, making for a total of 21 possible points. As the review quiz had three different question types, two of which students had not previously experienced, an ANOVA was conducted for each section of the quiz as well as one to compare the final scores. There were no significant differences found between the groups in the final score or in when looking at each section in isolation:

- Section 1: $F(2, 99) = .589$, $p. = .557$
- Section 2: $F(2, 99) = 1.902$, $p. = .155$
- Section 3: $F(2, 99) = 1.378$, $p. = .257$
- Total Score: $F(2, 99) = 1.480$, $p. = .233$

Means (standard deviations) from Sections 1 through 3 in addition to the total score of the Review Quiz are shown in Table 3.

Table 3 Mean scores (SD) for Review Quiz (n = 102)

Group	Section 1	Section 2	Section 3	Total Score
Paper 1 (n=29)	7.10 (2.30)	1.03 (0.91)	3.79 (1.68)	11.93 (4.10)
Paper_2 (n=37)	6.51 (2.73)	0.68 (0.88)	3.16 (1.79)	10.35 (4.44)
Smartphone (n=36)	6.44 (2.79)	0.64 (0.87)	3.14 (1.80)	10.22 (4.48)

Discussion

Regarding RQ1, the scores of the experimental group displayed no apparent advantage or disadvantage due to their use of smartphones to take their vocabulary quizzes on the LMS Canvas. Though this group had the lowest mean score in both of the quizzes, the differences between groups was not statistically significant. This lack of test mode effect has been seen in other studies (Čandrlić et al., 2014; Noyes et al., 2004; Vanpatten et al., 2015).

As for RQ2, though the experimental group had the lowest mean scores in every category, again, the differences were not statistically significant. It appears that if there was a testing effect, the test mode did not have enough of an impact to differentiate the experimental and control groups. These results are in contrast to other studies in foreign language vocabulary testing which found that students who had written their previous answers had performed better than those who took digital tests (Longcamp et al., 2006; McDaniel et al., 2007).

Regarding the different sections of the review quiz, Section 1 was the standard format of the regular quizzes and could be answered with the help of the included Japanese translation. Section 2 was added as an attempt to differentiate between working from the supplied Japanese or memorizing the English phrases from the vocabulary textbook. Due to the many possibilities that could fill any of those phrase blanks, students could only be expected to answer correctly if they had memorized the English phrase. As example, item 14 was“review a (p_____)”. The phrase in the textbook used“proposal”, but without the Japanese translation to indicate what word filled the blank, many other words were possible: plan, paragraph, page, picture, etc. Answers that did not match the word used in the example phrase were not counted as correct. Section 3 was designed to better check if students had learned the vocabulary words and had not just memorized them. The quiz design seemed to function as intended with a relatively high effect size between sections 1 and 3 (r=.594, p. <.001) and a relatively low effect size between sections 2 and 3 (r=.226, p.=.023) which, as an aside to this study, indicated that students were learning the vocabulary and not just memorizing it.

Implications

While this paper looked at testing mode as it related to cognitive load, another factor that should be considered is student perception (Dashtestani, 2015). During the time frame that this study took place, several students in the control groups expressed concerns to their teachers that the experimental group had an advantage on the vocabulary quizzes. They felt that due to predictive text and automatic spell checks that are common with modern smartphones, students who used smartphones would automatically perform better. This seemed like a valid concern, however, when doing post-quiz checks of the experimental group, regular misspellings were often found, indicating that spellcheck and autocorrect were not factors in student answers. That said, prior to the administration of Quiz 6, it was decided that if it became apparent that one group of students was being disadvantaged due to these different testing modes, the study would end, and the experimental and control groups would begin receiving quizzes in the same formats.

Another concern with shifting to online testing is consideration of the ease with which students are able to cheat. Using computers or mobile devices, students have the capability to easily jump back and forth between the testing page and "cheat sheets" without the teacher noticing unusual behavior. The prevalence of cheating in online assessments has been reported in the literature (Sullivan, 2016) and an explanation for this lack of academic integrity was offered by Davies and West (2014), who reported that undergraduate students found it easier to cheat when online and that they might have a different perception of what constitutes cheating due to their experience with online sharing and collaboration.

In fact, with the experimental group in this study, cheating was found to be a problem in the first two administrations of the Canvas quiz. While unknown prior to the implementation of these quizzes, the LMS that was used, Canvas, keeps a log of when/how long students leave the quiz page for another. When evidence indicating cheating was found after the first quiz, scores were changed to zeros and students were warned. In the second quiz, a few more students cheated and were also given zeros. After the changed grades and a second classroom discussion on academic integrity, there were no incidents of cheating on the remaining quizzes. In their program, Vanpatten et al. (2015) addressed the concern about cheating by reframing the way the quizzes were viewed. Instead of having students think of them as high stakes "quizzes," they were instead labeled as end-of-unit review activities.

Limitations

The main limitation of this study is the lack of generalizability due to nature of the quizzes and the

specificity of the context within which they were given. While it could be expected that some digital competencies are required to take a quiz online, with these quizzes, students had 10 minutes to input 10 vocabulary words. Under such conditions, the quizzes were a matter of recall and simple input and as indicated by the statistical analyses, there was no test mode effect in evidence. However, had the questions been longer and more intricate or had the answers required short essays, the use of smartphones vs. paper-based testing modes might have made a difference because transferring thoughts to text with digital keyboards on smartphone screens requires a different set of skills than putting pen to paper. Another limitation was with the design of the review quiz. Because each section had a different number of items, it is possible that the analyses comparing them were not as robust as they would have been had all sections had an equal number of items.

Conclusion

The impetus of this study was to see if there were any observable differences regarding student scores on twice-weekly vocabulary quizzes that could be attributed to testing mode. It arose as a natural artifact when one teacher, for reasons of efficiency, began incorporating online quizzes with one section of a required course when teachers of other sections were using paper-based quizzes. While no effect was observed, the quizzes themselves were extremely basic with seemingly very low cognitive load. If the quizzes had been more intrinsically taxing, it is possible the results would have been different. It should be kept in mind that other studies have shown that testing mode can have an impact on student results or on perceptions of difficulty and fairness (Čandrlić et al., 2014; Clariana & Wallace, 2002; Noyes et al., 2004). Even if it was not the case in this study, it is possible that different types of tests would expose statistically significant differences depending on the testing mode.

In practice, when integrating technology, especially for testing purposes, student familiarity with the technology and the potential for an overabundance of extraneous cognitive load needs to be taken into account. Even when all else is equal, a timed evaluation using a digital assessment might result in scores that are at least in part the result of students' ability to input text relatively quickly and accurately.

As the variables of measures used in student evaluation are tethered to the local context, researchers and practitioners should keep in mind the potential for the test mode to have some effect, either positive or negative, and as a result obscure the primary reason for the test: evaluating learning outcomes. This potential interference highlights the need to test for the test mode effect, especially in a coordinated program where, in order to ensure fairness, all students should be evaluated with the measures.

References

Aldunate, R., & Nussbaum, M. (2013). Teacher adoption of technology. *Computers in Human Behavior, 29*(3), 519–524. https://doi.org/10.1016/j.chb.2012.10.017

Antonenko, P. D., & Niederhauser, D. S. (2010). The influence of leads on cognitive load and learning in a hypertext environment. *Computers in Human Behavior, 26*(2), 140–150. https://doi.org/10.1016/j.chb.2009.10.014

Čandrlić, S., Katić, M. A., & Dlab, M. H. (2014). Online vs. Paper-based testing: A comparison of test results. *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2014 - Proceedings, May*, 657–662. https://doi.org/10.1109/MIPRO.2014.6859649

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology, 33*(5), 593–602. https://doi.org/10.1111/1467-8535.00294

Clark, R. E. (1983). Reconsidering Research on Learning from Media. *Review of Educational Research, 53*, 445–459.

Clark, R. E. (1994). Media Will Never Influence Learning. *Educational Technology Research and Development, 42*(2), 21–29.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed method*. https://doi.org/10.1007/s13398-014-0173-7.2

Dashtestani, R. (2015). Examining the Use of Web-Based Tests for Testing Academic Vocabulary in EAP Instruction. *Teaching English with Technology, 15*(1), 48–61. http://ezproxy.uow.edu.au/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1140573&site=ehost-live

Davies, R. S., & West, R. E. (2014). Technology Integration in Schools. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology: Fourth Edition* (pp. 841–853). Springer. https://doi.org/10.1007/978-1-4614-3185-5_68

Dukic, Z., Chiu, D. K. W., & Lo, P. (2015). How useful are smartphones for learning? Perceptions and practices of Library and Information Science students from Hong Kong and Japan. *Library Hi Tech, 33*(4), 545–561. https://doi.org/10.1108/LHT-02-2015-0015

Ertmer, P. A., & Ottenbreit-Leftwich, A. T. (2010). Teacher Technology Change How Knowledge, Confidence, Beliefs and Culture Intersect. *Journal of Resaerch on Technology in Education, 42*(3), 255–284. https://doi.org/10.1080/15391523.2010.10782551

Ertmer, P. A., Ottenbreit-Leftwich, A. T., Sadik, O., Sendurur, E., & Sendurur, P. (2012). Teacher beliefs and technology integration practices: A critical relationship. *Computers and Education, 59*(2), 423–435. https://doi.org/10.1016/j.compedu.2012.02.001

Gitsaki, C., & Robby, M. A. (2014). Post-Secondary Students Using the iPad to Learn English: *International Journal of Mobile and Blended Learning, 6*(4), 53–74. https://doi.org/10.4018/ijmbl.2014100104

Hastings, N. B., & Tracey, M. W. (2004). Does media affect learning: where are we now? *TechTrends, 49*, 28–30. https://doi.org/10.1007/BF02773968

Kozma, R. B. (1991). Learning with Media. *Review of Educational Research, 61*(2), 179–211. https://doi.org/10.3102/00346543061002179

Liu, S. H. (2011). Factors related to pedagogical beliefs of teachers and technology integration. *Computers and Education, 56*(4), 1012–1022. https://doi.org/10.1016/j.compedu.2010.12.001

Longcamp, M., Boucard, C., Gilhodes, J. C., & Velay, J. L. (2006). Remembering the orientation of newly learned characters depends on the associated writing knowledge: A comparison between handwriting and typing. *Human Movement Science, 25*(4–5), 646–656. https://doi.org/10.1016/j.humov.2006.07.007

Mangen, A., Anda, L. G., Oxborough, G. H., & Brønnick, K. (2015). Handwriting versus keyboard writing : Effect on word recall Han dwritin ng vers us Key yboard Writing g : Effec ct on Word Re ecall. *Journal of Writing Research, 7*(2), 227–247.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*(4–5), 494–513. https://doi.org/10.1080/09541440701326154

McKnight, K., O'Malley, K., Ruzic, R., Horsley, M. K., Franey, J. J., & Bassett, K. (2016). Teaching in a Digital Age: How Educators Use Technology to Improve Student Learning. *Journal of Research on Technology in Education, 48*(3), 194–211. https://doi.org/10.1080/15391523.2016.1175856

Noyes, J., Garland, K., & Robbins, L. (2004). Paper-based versus computer-based assessment: Is workload another test mode effect? *British Journal of Educational Technology, 35*(1), 111–113. https://doi.org/10.1111/j.1467-8535.2004.00373.x

Oblinger, D., & Hawkins, B. (2006). IT myths: The myth about no significant difference. *Educause Review, 41*(6), 14–15. http://www.educause.edu/ero/article/myth-about-no-significant-difference

Richey, R. C., & Klein, J. D. (2014). Design and Research Development. In *Handbook of Research on Educational Communications and Technology* (pp. 141–150). https://doi.org/10.1007/978-1-4614-3185-5

Spires, M., & Bartlett, H. (2012). Digital Literacies and Learning : Designing a Path Forward. *Friday Institute White Paper Series, No. 5*(June), 1–24. www.fi.ncsu.edu/whitepapers

Spotts, T. H. (1999). Discriminating factors in faculty use of instructional technology in higher education. *Educational Technology and Society, 2*(4), 139–150.

Straub, E. T. (2009). Understanding Technology Adoption: Theory and Future Directions for Informal Learning. *Review of Educational Research, 79*(2), 625–649. https://doi.org/10.3102/0034654308325896

Sullivan, D. P. (2016). An Integrated Approach to Preempt Cheating on Asynchronous , Objective , Online Assessments in Graduate Business Classes. *Online Learning, 20*(3), 195–209.

Sweller, J., Van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive Architecture and Instructional Design. *Educational Psychology Review, 10*(3), 251–296.

Vanpatten, B., Trego, D., & Hopkins, W. P. (2015). In-Class vs. Online Testing in University-Level Language Courses: A Research Report. *Foreign Language Annals, 48*(4), 659–668. https://doi.org/10.1111/flan.12160

Appendix A

Vocabulary Quiz 6

TOEIC 単語テスト（6）範囲　101～120

空所に適切な英語を入れなさい。解答は解答欄にすること

1. open a new (b　)（新しい支店をオープンする）
2. (p　) leave（有給休暇）
3. in (o　) condition（元の状態で）
4. a (r　) increase（家賃の値上げ）
5. an art (e　)（美術展）
6. a (l　) company（トップの会社）
7. lead an (o　)（組織を率いる）
8. for a (l　) time（期間限定で）
9. a normal (p　)（通常の手続き）
10. employee (b　)（従業員の福利厚生）

Appendix B

Review Quiz

1. Please (r　) to that.（それを参照してください）
2. the sales (d　)（営業部）
3. (r　) a report（報告書に目を通す）
4. (d　)s of a plan（計画の詳細）
5. a (c　) store（衣料品店）
6. a successful (c　)（合格した候補者）
7. public (t　)（公共の交通機関）
8. (i　) a new line（新たなラインを導入する）
9. local (r　)s（地元の住民）
10. (r　) parts（交換部品）

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

11. a large (c　) room
12. (a　) the winners
13. a Q&A (s　)
14. review a (p　)
15. returns (p　)

11	
12	
13	
14	
15	

16. Who, what, when, where, and why are the (d　) you should include in your story.
17. Softbank will (r　) a broken iPhone screen for about 5,000 yen.
18. Uniqlo is famous for selling fashionable and cheap (c　).
19. The (m　) fee for this Konami Sports Club is 8,750 yen a month.
20. All of the office (s　) are in room 411, so if you need things like paper or pens, you can get them there.
21. Last year, I lived in Takarazuka, but now I am a (r　) of Ibaraki city.

16	
17	
18	
19	
20	
21	