frontiers in Education

ORIGINAL RESEARCH
published: 09 July 2018
doi: 10.3389/feduc.2018.00051

A Comparison of the Effectiveness of Two Computer-Based Learning Aids

Piers D. L. Howe[1]*, *Jason M. Lodge*[2] *and Meredith McKague*[1]

[1] *Melbourne School of Psychological Sciences, University of Melbourne, Melbourne, VIC, Australia,* [2] *School of Education, University of Queensland, Brisbane, QLD, Australia*

Instructors are increasingly using computer-based educational technologies to augment their courses. As answering quizzes has been shown to be one of the most effective learning strategies, a growing number of computer-based learning aids use quizzing. So which of these learning aids should instructors recommend to their students? These learning aids typically either present the student with a number of potential answers and require that they recognize the correct answer (i.e., a multiple-choice quiz) or else they might require the student to recall the answer without assistance (i.e., a free recall quiz). Numerous lab-based studies have shown that recall-based quizzes promote more learning and result in higher performance in a subsequent exam/test than recognition-based quizzes. In the present study, we investigated to what extent this finding holds in an actual university setting with two commercially-available learning aids. We found that while both types of learning aid proved to be effective, we could find no evidence that the recall-based learning aid was more effective than the recognition-based learning aid. In light of this, we discuss possible reasons why the laboratory findings did not readily translate to an actual university setting and make practical recommendations for what sort of computer-based learning aid instructors should incorporate into their university courses.

Keywords: educational psychology, revision aid, Cram, PeerWise, computer-assisted learning, learning

OPEN ACCESS

Edited by:
Meryem Yilmaz Soylu,
University of Nebraska-Lincoln, United States

Reviewed by:
Larry Robert Medsker,
George Washington University, United States
Jeffrey K. Smith,
University of Otago, New Zealand

***Correspondence:**
Piers D. L. Howe
pdhowe@unimelb.edu.au

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Education

Received: *26 March 2018*
Accepted: *13 June 2018*
Published: *09 July 2018*

Citation:
Howe PDL, Lodge JM and McKague M (2018) A Comparison of the Effectiveness of Two Computer-Based Learning Aids. Front. Educ. 3:51. doi: 10.3389/feduc.2018.00051

INTRODUCTION

The purpose of educational technology is to facilitate learning (Robinson et al., 2016). Computer-based learning aids are becoming increasingly popular (Liaw et al., 2007). Many of these learning aids employ quizzes, as quizzing has been shown to be one of the most effective teaching strategies (Dunlosky et al., 2013). In particular, it has been shown that taking a quiz improves long-term retention more than spending an equivalent amount of time restudying the material (Roediger and Karpicke, 2006; Karpicke and Roediger, 2008). This effect is especially pronounced if feedback is given (Kang et al., 2007; Roediger and Butler, 2011). This *testing effect*, whereby engaging in effortful retrieval improves memory for the material studied, has been widely demonstrated in laboratory settings (Roediger and Butler, 2011; Roediger et al., 2011), but studies in actual educational contexts have produced mixed results (Karpicke et al., 2014; Nguyen and McDaniel, 2015; Trumbo et al., 2016).

Quizzes can be broadly classified as being either recognition-based or recall-based (Carpenter and DeLosh, 2006; Dunlosky et al., 2013). In recognition-based quizzes, the participant is presented with multiple potential answers and is required to identify the correct one. Multiple choice questions are a popular example of a recognition-based quiz. Conversely, in recall-based quizzes,

no options are presented and the student is instead required to recall the answer without assistance (Dunlosky et al., 2013). Previous laboratory studies have indicated that recall-based quizzes result in greater learning and greater retention of the tested material than recognition-based quizzes, as evidenced by increased scores on a subsequent exam or quiz (Bjork and Whitten, 1974; Glover, 1989; Carpenter and DeLosh, 2006; McDaniel et al., 2013). Presumably, this is because recall quizzes, often being more difficult (McDaniel et al., 2007), provide more potential for elaborative processing, which results in more learning (Carpenter and DeLosh, 2006).

It is unclear to what extent these laboratory-based findings apply to actual higher education courses (Karpicke et al., 2014). In these previous lab-based studies the final exam was either the same as or at least was very similar to the revision quiz (Bjork and Whitten, 1974; Glover, 1989; Carpenter and DeLosh, 2006; McDaniel et al., 2013). This meant that by taking the revision quiz, the students could obtain a good idea of the content of the final exam. For many university courses, this would not be acceptable as this could cause students to ignore most of the course and focus only on the material that would appear on the final exam (Wooldridge et al., 2014). When the questions in the final exam do not closely match the questions in the quizzes, the benefits of quizzes are typically reduced and sometimes eliminated (Rohrer et al., 2010; Wooldridge et al., 2014; Nguyen and McDaniel, 2015). Consequently, we do not know whether recall-based quizzes will prove to be superior to recognition-based quizzes in an actual higher education setting or indeed whether either type of quiz will aid learning at all (Nguyen and McDaniel, 2015). The issue is further complicated by the fact that university courses typically teach material that is interrelated, as opposed to comprising merely a list of independent facts that can be learned in isolation from each other. It is know that the testing effect is reduced and may even be eliminated for material that is interrelated (Van Gog and Sweller, 2015), making it hard to predict the relative effectiveness of these two types of quizzes in an actual university setting.

A secondary issue regards the way the students interacted with the quizzes in these previous studies. In the previous laboratory-based studies, students were typically required to answer each quiz question (Bjork and Whitten, 1974; Glover, 1989; Carpenter and DeLosh, 2006; McDaniel et al., 2013). Conversely, in an actual classroom setting it is often hard to enforce substantial participation with the quizzes. The degree to which the students choose to engage with the learning aids will largely determine their effectiveness (Rawson and Dunlosky, 2011). Consequently, without knowing to what degree students will actually engage with these learning aids, it is impossible to predict their relative effectiveness.

In summary, a number of previous laboratory-based studies have indicated that recall-based quizzes should be more effective than recognition-based quizzes in helping students learn course material. However, it is unclear to what extent these findings will translate to an actual university setting using commercially available learning aids that are commonly used by students. The purpose of our study was to address this gap in the literature by comparing the effectiveness of two popular learning aids in a university setting. One learning aid was chosen because it employed recognition-based quizzes, while the other was chosen because it employed recall-based quizzes. Our expectation was that because recall-based quizzes are more difficult, so require more effort, they will result in more learning, and in higher performance in a subsequent test (McDaniel et al., 2007). We therefore had two specific hypotheses. The first was that both learning aids would prove to be effective in that participation with them would be correlated with higher marks in the final exam, even when potential confounding factors are discounted. Our second hypothesis was that this correlation between participation with a learning aid and higher exam performance would be larger for the recall-based learning aid than the recognition-based learning aid, reflecting the previous laboratory findings (Bjork and Whitten, 1974; Glover, 1989; Carpenter and DeLosh, 2006; McDaniel et al., 2013). By testing these two hypotheses, we hoped to gain some insight into the relevant merits of recognition-based vs. recall-based quizzes, so that we could advise university instructors as to which type of learning aid they should choose for their university courses.

METHODS

The Two Computer-Based Learning Aids

Since the purpose of our study was to compare the effectiveness of a retrieval-based learning aid vs. a recall-based learning aid in an actual university setting, we opted to use commercially available educational technologies that students would actually engage with, as opposed to home-grown learning aids designed for research purposes, as have been used in previous studies (Bjork and Whitten, 1974; Glover, 1989; Carpenter and DeLosh, 2006; McDaniel et al., 2013). Using commercially available educational technologies created difficulties as to how participation could be equivalently measured for the two learning aids. We discuss how this issue was addressed later.

For the recall-based learning aid, we chose to use a flashcard system, as flashcards have been a consistently popular learning aid among students (Dunlosky et al., 2013). We further chose to use electronic vs. physical flashcards as it was unfeasible to provide real flashcards for all the students in the course. A second advantage of electronic flashcards is that we could monitor their usage. An informal poll of a number of students revealed that there were several electronic flashcard systems in common usage among the student population. We chose to use the Cram system (www.cram.com) because it was already popular with the students, it was free and it was compatible with all the computer systems that the students were likely to use. There were a number of other systems that were similar to Cram (e.g., Quizlet) and were also popular with the students. Had we used one of these other systems, we expect that we would have obtained essentially the same results.

For the recognition-based learning aid, we chose to use multiple choice quizzes as these have often been used in prior studies (McDaniel et al., 2013). A number of learning aids that use multiple choice quizzes, but we chose to use the PeerWise System (Denny et al., 2008a,b; Rhodes, 2013). This is a particularly popular internet-based learning aid that is also free to use and readily accessible by students.

The Course

The two computer-based memorization aids were introduced into a large second year Biological Psychology course at a large Australian university in 2016. This class lasted one semester, comprising 12 weeks. Each week students received two 1-h lectures. During the course, students were required to write two essay-based assignments, due in weeks 4 and 8 and sit a multiple choice final exam that occurred 2 weeks after the end of the course. This exam covered the entire course, but in a relatively superficial manner that primarily tested the students' ability to memorize simple facts, such as what certain acronyms stood for or the spatial resolution of particular imaging techniques. As such, the questions were very similar in format and style to the PeerWise questions discussed below. Thus, students needed to focus more on memorizing basic factual information than analyzing or evaluating it. As such, it was an ideal course for testing computer-based memorization aids, which are designed to assist in learning this type of information.

Procedure

This study was carried out in accordance with the recommendations of the Melbourne School of Psychological Sciences (MSPS) Human Ethics Advisory Group. The protocol was approved by MSPS Human Ethics Advisory Group, approval number 1647103. All participants gave written informed consent in accordance with the Declaration of Helsinki. It was an explicit requirement of our ethics approval that all students be able to access both learning aids. This constraint excluded a randomized design, in which students would have been randomly allocated to one of the two learning aids or to neither of them (i.e., a control group). This would have been our preferred procedure. Instead, we opted to control for confounding factors using partial correlations and regressions, as discussed below.

Cram Electronic Flashcards

For each lecture, a set of electronic flashcards were prepared by a pair of tutors. Thus, 24 sets of flashcards were produced in total. Each set contained on average 20.8 flashcards, range 16–28 flashcards. These tutors did not have any knowledge of the final exam. In this way, we ensured that the flashcards did not give any hints as to what would be on the final exam. The flashcards were checked for accuracy by the faculty member who gave the lecture. These flashcards were made available to the students within a week of the lecture being given. They were designed by the tutors to systematically cover the most important points of the lecture. Students were required to electronically sign a statement saying that they had downloaded and used at least one set of flashcards. Participation beyond this level was not enforced. Two example flashcard questions and answers are listed below:

Question: What is EEG?
Answer: Electroencephalography (EEG) is a method of detecting neural activity by placing electrodes on the scalp.
Question: What is retrograde amnesia?
Answer: Impairment for memories that were created before injury.

Students were free to use these flashcards in any way they saw fit. Although they were encouraged to write down their answer before viewing the official answer, this was not enforced, and we have no way of determining how often they in fact did this. As discussed later, face-to-face interviews with students suggest that they did not always do this.

PeerWise

A PeerWise website was created for this course (Denny et al., 2008a,b; Rhodes, 2013). On this website, students could create, share and answer multiple choice questions created by other students. Written instructions were given to students as to how to do this. To ensure participation, each student was required to create at least one multiple choice question. Participation beyond this level was not enforced. While the PeerWise questions had a similar style and format to those question on the final exam, as the students that wrote the PeerWise questions had no knowledge of the questions on the final exam, the content was necessarily different. An example PeerWise question is listed below:

For a chemical to qualify as a neurotransmitter it must:

A. be present in the presynaptic neuron, released from the presynaptic neuron after it fires, cause the postsynaptic neuron to fire
B. cause the ion channel to open on the presynaptic neuron, released from the presynaptic neuron after it fires, cause an ion channel to open on the postsynaptic neuron
C. be synthesized in the presynaptic neuron, have receptors sensitive to it on the postsynaptic neuron, cause an ion channel to open on the presynaptic neuron
D. be present in the presynaptic neuron, released from the presynaptic neuron after it fires, have receptors sensitive to it on the postsynaptic neuron
E. be synthesized in the presynaptic neuron, released from the presynaptic neuron after it fires, cause the postsynaptic neuron to fire

(Correct answer is D).

Measures

Survey

In the week after the course concluded, students were administered a 7-question survey asking for their subjective impressions regarding these two learning aids. The survey was as follows:

Considering only the course *Biological Psychology*:

Q1: Please estimate how many hours you spent using PeerWise:
Q2: Please estimate how many hours you spent using Cram flashcards:
Q3: Which did you find more useful for your revision*? (Choice: "PeerWise" or "Cram")
Q4: To what extent did using PeerWise affect the total number of hours you spent revising for the exam? (Scale 1–5: 1 = Significantly reduced, 2 = Slightly reduced, 3 = No change, 4 = Slightly increased, 5 = Significantly increased)

Q5: To what extent did using Cram flashcards affect the total number of hours you spent revising for the exam? (Scale 1–5: 1 = Significantly reduced, 2 = Slightly reduced, 3 = No change, 4 = Slightly increased, 5 = Significantly increased)
Q6: To what extent did using PeerWise make you feel confident that you understood the material covered in this course? (Scale 1–11: 1 = Not at all, 11 = Completely)
Q7: To what extent did using Cram flashcards make you feel confident that you understood the material covered in this course? (Scale 1–11: 1 = Not at all, 11 = Completely)

*In Australia, the term *revision* is commonly used to mean to review material to increase one's knowledge and understanding of it.

Objective Measures

In addition, to these subjective impressions, we collected objective data. Because the Cram flashcard system was designed primarily for teaching purposes, there was no built-in facility for providing detailed information on student usage. For this reason, we operationalized the student's (objective) usage of the flashcards as the number of days where the student accessed at least one set of flashcards. For the sake of comparison, we defined usage of PeerWise in an analogous manner. In addition, for each student we recorded their mark for each assignment and their exam score.

Focus Groups

We also organized two focus groups to discuss the results with the students. These groups used a semi-structured format, led by the moderator, based around the following questions:

1) Do you believe PeerWise or Cram was the better revision aid? Why?
2) To what extent did using PeerWise affect the total number of hours you spent revising for the exam? Why?
3) To what extent did using the Cram flashcards affect the total number of hours you spent revising for the exam? Why?
4) To what extent did using PeerWise make you feel confident that you understood the material covered in this course? Why?
5) To what extent did using the Cram flashcards make you feel confident that you understood the material covered in this course? Why?
6) Do you think the number of hours spent on PeerWise and Cram would correlate with the final exam score? Why not?
7) Did using Cram/PeerWise make you feel more confident that you understood the material. Do you think that feeling confident would be associated with scoring higher on the final exam?

RESULTS

Survey

Six hundred ninety-nine students completed the course and sat the final exam. Of these, 262 completed the survey. The key findings are as follows:

Preference for Cram

On average, students felt that Cram was more helpful than PeerWise [$\chi^2(1, N = 262) = 17.6$, $p < 0.001$] with 63% preferring Cram to PeerWise. Consistent with this preference, students reported spending significantly more time on Cram than on Peerwise [$t_{(261)} = 5.17$, $p < 0.001$, $r^2 = 0.09$], reporting spending on average 10.6 (SEM = 0.75) hours on Cram and 6.15 (SEM = 0.60) hours on PeerWise.

Cram Made Students Feel More Confident That They Understood the Material

The use of Cram and PeerWise made people somewhat confident that they understood the material covered in the course. The degree of confidence induced by the learning aid was 7.48 (SEM = 0.17) for Cram and 6.35 (SEM = 0.17) for PeerWise. The rating was significantly higher for Cram than for PeerWise [$t_{(261)} = 4.70$, $p < 0.001$, $r^2 = 0.08$].

Confidence Negatively Correlated With Exam Performance

The more confident Cram made them feel, the *worse* they did on the final exam [$r_{(260)} = -0.129$, $p = 0.038$]. A similar trend was observed for PeerWise, but this did not reach significance [$r_{(260)} = -0.100$, $p = 0.11$]. However, when performance on the two assignments was discounted, neither the PeerWise nor the Cram partial correlation was significant [PeerWise: $r_{(260)} = 0.071$, $p = 0.25$; Cram: $r_{(260)} = -0.043$, $p = 0.48$].

For Cram, Confidence Negatively Correlated With Performance in the Assignments

The degree to which Cram made students feel confident that they understood the material covered in this course was negatively correlated with their assignment scores [Assignment 1: $r_{(260)} = -0.129$, $p = 0.037$; Assignment 2: $r_{(260)} = -0.149$, $p = 0.016$]. For PeerWise, only the first correlation was significant and that was in the opposite direction, [Assignment 1: $r_{(260)} = 0.134$, $p = 0.031$; Assignment 2: $r_{(260)} = 0.024$, $p = 0.70$].

Students Who Preferred Cram Did Worse on the Final Exam

Students who preferred Cram tended to do worse on the final exam than those who preferred PeerWise [$t_{(260)} = 2.08$, $p = 0.038$, $r = 0.016$].

Neither Learning Aid Appeared to Be Effective According to These Subjective Measures

There was no significant correlation between the final exam score and the self-reported number of hours spent using either PeerWise [$r_{(260)} = 0.045$, $p = 0.47$] or Cram [$r_{(260)} = -0.009$, $p = 0.88$].

Objective Measurements of Usage

Given that previous studies have found a significant correlation between usage of PeerWise and final exam score (Hardy et al., 2014; McQueen et al., 2014), we were surprised at the lack of correlation between the self-reported number of hours spent using PeerWise and the final exam score. It is possible that the students' estimates of their own usage of these learning

aids may be so noisy that they hide the correlation. This is why we chose to additionally measure usage in an objective fashion. Because with Cram we could not determine how many questions each student answered, we instead measured usage as the number of days students were objectively recorded as accessing each learning aid. We discuss alternative measures of usage below. We were also concerned that our findings might be confounded by student ability. Specifically, we were concerned that the more able students may tend to use the learning aids more. As these students would tend to perform better in the final exam, this might give a false impression of the effectiveness of the learning aids. We, therefore, decided to factor out student ability. We operationalized student ability as their individual scores on the two assignments (i.e., we controlled for both scores simultaneously). Finally, so as to ensure that we measured the unique contribution of each learning aid, we controlled for the time spent using the other learning aid. Thus, when we performed the partial correlations to test for a relationship between the number of days spent using a learning aid and the final exam score, we controlled for the marks in both assignments and the time spent with the other learning aid.

Six hundred and ninety-one students completed both assignments and the exam. A student could interact with PeerWise by either authoring, answering or commenting on a question, though in practice in 97.8% of the interactions with PeerWise a student answered a question, in 1.7% of the interactions a student authored a question and in only 0.6% of the interactions did the student comment on a question. When a student interacted with Cram, their only option was to answer a question.

Both participation with Cram and participation with PeerWise uniquely predicted final exam scores. The degree of participation with PeerWise was still significantly correlated with final exam score, even when both student ability and the degree of participation with Cram was controlled for [$r_{(689)} = 0.170$, $p < 0.001$]. Similarly, the degree of participation with Cram was still significantly correlated with the final exam score, even when both student ability and the degree of participation with PeerWise was controlled for [$r_{(689)} = 0.101$, $p = 0.006$]. Running a bootstrap analysis with replacement revealed no significant difference between these two correlations, $p = 0.18$ (Efron and Tibshirani, 1998). We ran a linear regression to estimate the degree to which participation with Cram and PeerWise affected final exam scores, accounting for performance on the first two assignments. To increase their final exam score by 5% (i.e., by one grade), on average students would have needed to have interacted with PeerWise on 10.1 days and with Cram on 24.3 days. Bootstrapping revealed this difference to be statistically significant, $p = 0.030$ (Efron and Tibshirani, 1998).

There was a significant correlation between usage of the two learning aids [$r_{(689)} = 0.41$, $p < 0.001$]. Despite this, student usage of the two learning aids also differed significantly, with students accessing PeerWise on fewer days than Cram [$t_{(690)} = 32.7$, $p < 0.001$, $r = 0.61$]. On average, students accessed PeerWise on 2.61 (SEM = 0.14) days. Conversely, on average students accessed Cram on 9.22 (SEM = 0.21) days. It was a course requirement that all students engage with both PeerWise and Cram on at least one occasion. Forty-nine percent of the students chose to engage with PeerWise more than once, and 99% chose to engage with Cram more than once.

Comparison With Other Objective Measures of Participation

For Cram, we could have also measured participation as the number of flashcard sets each student accessed. Students were allowed to access the same set multiple times and we recorded how many times they accessed each flashcard set. The correlation between this measure of participation and our actual measure of participation was $r = 0.89$ ($p < 0.001$), showing that the two measures were highly similar. Using this new measure of Cram participation, we got a very similar answer as before: the degree of participation with Cram was still significantly correlated with final exam score, even when both student ability and the degree of participation with PeerWise was controlled for [$r_{(689)} = 0.080$, $p = 0.036$]. A bootstrap analysis revealed that this correlation was not significantly different from the equivalent correlation calculated using the other measure of participation with Cram, $p = 0.18$ (Efron and Tibshirani, 1998).

We could have measured PeerWise participation as the total number of PeerWise questions that were answered, created or commented on by a given student. The correlation between this measure and our actual measure of participation was $r = 0.72$ ($p < 0.001$). Using this measure of participation we found that the degree of participation with PeerWise was still significantly correlated with final exam score, even when both student ability and the degree of participation with Cram was controlled for [$r_{(689)} = 0.197$, $p < 0.001$]. As before, a bootstrap analysis revealed that this correlation was not significantly different from the equivalent correlation calculated using the other measure of participation with PeerWise, $p = 0.32$ (Efron and Tibshirani, 1998).

Focus Groups

Two focus groups were held to discuss these results with the students. These were voluntary and held 2 weeks after the final exam. In total, 12 students attended. These students were equally split in preference for Cram vs. PeerWise and there was no consensus as to which was the better learning aid. They reported finding both to be highly engaging. They especially liked the fact that the questions on PeerWise were in a very similar format to those on the final exam. They also liked the fact that the Cram electronic flashcards were compatible with their mobile phones, permitting them to study whenever it was convenient (e.g., on the train or on the tram). Consistent with the objective measurements, they didn't feel that either learning aid substantially increased the amount of time they spent revising. They reported that the reason for this was that they used these learning aids instead of revising in their usual manner (e.g., making notes). Their major concern with the Cram electronic flashcards was that their content and format often closely resembled and duplicated the lecture notes (on which the flashcards were directly based). Their major concern with PeerWise was that they felt that a significant fraction of the questions were of low quality and poorly written. Consistent with

this concern, a review of the questions on PeerWise found that only 18% received a rating of at least "good," 25% received a rating of "poor" or "very poor," while the remainder received a rating of "fair." Despite this, they agreed that using both Cram and PeerWise made them feel more confident about the material covered in the course, consistent with the survey results. They were unable to agree on which one made them feel more confident. They expected that the more confident a learning aid made them the better they would do on the final exam. They were surprised when informed that the converse occurred. They were not surprised that the number of self-reported hours spent with the learning aids did not correlate with the final exam performance. They explained this by suggesting that these estimates were probably very unreliable. When asked why Cram was not more effective than PeerWise they reported that with Cram you could go "easy on yourself." By this they meant that when using Cram students were not forced to write down an answer, which could mask the fact that they didn't know the answer. Conversely, in PeerWise they were forced to choose an answer and then received immediate feedback on their choice. While they found that this was more confronting, they reported that this helped them form a more accurate assessment of their ability.

DISCUSSION

Computer-based tools and applications are becoming increasingly popular and widely used in higher education (Liaw et al., 2007). As such, we need to evaluate their effectiveness in an actual higher education setting (Lodge and Horvath, 2017). The purpose of the current paper was to compare the effectiveness of two computer-based learning aids that employed quizzes to help students learn course material. These particular computer-based aids were chosen because they used different types of quizzes. One used recall-based quizzes, while the other used recognition-based quizzes. Previous lab-based studies have shown that both recall-based quizzes and recognition-based quizzes are effective at promoting learning but recall-based quizzes result in more learning and higher performance in a subsequent test than recognition-based quizzes (Bjork and Whitten, 1974; Glover, 1989; Carpenter and DeLosh, 2006; McDaniel et al., 2013). Based on these findings, our expectation was that both Cram and PeerWise would be effective learning aids but Cram would be more effective than PeerWise as Cram employed recall-based quizzes whereas PeerWise employed recognition-based quizzes. Consistent with our first hypothesis, we found that participation with either learning aid was significantly correlated with improved exam performance, even when student ability and participation with the other learning aid was discounted. Contrary to our second hypothesis, we found no significant difference between the correlation between participation with Cram and the final exam and the correlation between participation with PeerWise and the final exam. Indeed, if anything, the correlation between participation with a learning aid and the final exam performance was less strong for Cram than for PeerWise. To quantify the effectiveness of each learning aid, we ran a linear regression. From this we determined that to increase their exam score by 5%, students need to access Cram on 24.3 occasions but PeerWise only on 10.1 occasions, a statistically significant difference. While this may indicate that PeerWise is more effective than Cram, this might be in part due to PeerWise being used less by students than Cram. Presumably, the more a learning aid is used, the less effective it becomes, so its average effectiveness decreases. It might also be that because Cram did not force students to respond whereas PeerWise did, PeerWise may have been genuinely more effective as a learning aid. Regardless of whether PeerWise is genuinely superior to Cram as a learning aid, we can say that our results certainly show no evidence that Cram is superior to PeerWise, contrary to our initial expectations.

One potential explanation for this surprising finding could be the format of our final exam. The final exam was entirely multiple choice, which is the same format employed by PeerWise but not the same format as used by Cram. However, a previous study has shown that, in the lab, recall-based learning quizzes still result in higher final exam performance than recognition-based quizzes even when the final exam was multiple choice (McDaniel et al., 2007). This would indicate that the format of the final exam was probably not the main reason why Cram failed to outperform PeerWise.

Another potential reason could be that Cram engendered a false sense of confidence and this decreased its effectiveness relative to PeerWise. Consistent with this, students reported that Cram made them feel more confident than PeerWise. We also found that the more confident a learning aid made a student feel, the worse they did on the final exam. However, it is possible that this correlation between being confident and doing worse on the final exam does not represent a causal relationship but rather is caused by both phenomena having a common cause (Aldrich, 1995). In particular, we believe that the common cause is likely to be student ability. Carpenter et al. (2016) have shown that the weaker students tend to be the most overconfident. Since we found that the weaker students also preferred Cram, this could explain our finding that Cram engendered more confidence than PeerWise. We tested this potential explanation by factoring out student ability using partial correlations. When student ability was factored out, there was no longer any correlation between how confident the learning aids made the students feel and final exam performance.

Another potential reason why Cram did not outperform PeerWise could be that students found Cram less engaging that PeerWise, so used it less. However, this appears not to be the case. The majority of the students preferred Cram and both subjective and objective measures indicated that they used it significantly more often than PeerWise.

We suggest that the most likely reason why Cram did not outperform PeerWise was because the quality of the feedback with Cram was less. With PeerWise, students were told whether or not they got the answer correct. Conversely, with Cram they were instead shown the correct answer and, on this basis, needed to judge how accurate their own answer was. While students reported finding this less confronting, it is possible that this reduced their learning. In the lab-based studies that employed recall-based quizzes, students were required to supply a specific answer, often a specific word or phrase (Bjork and Whitten, 1974; Glover, 1989; Carpenter and DeLosh, 2006; McDaniel et al.,

2013). When they were shown the correct answer, it was easy for them to judge whether they had gotten the answer correct or not. Conversely, with the Cram flashcards, the required answers were often longer and could potentially be phrased in different ways, making it harder for students to judge how correct their answer was. This would presumably reduce the quality of feedback that the students would receive and might give the students the impression that they know more than they do. Although it is not essential, feedback does improve learning via quizzes (Kang et al., 2007; Roediger and Butler, 2011), so presumably the higher the quality of the feedback, the more learning will occur. If Cram did indeed provide lower quality feedback than PeerWise, this could explain why it failed to be more effective as a learning aid than PeerWise.

Implications and Limitations

For ethical reasons, we were required to provide all students with access to both learning aids, so could not perform a randomized control experimental design. Instead, we performed a correlational analysis. This means that we cannot determine with certainty what the cause and effect relationships are. A correlation between two variables can be caused by both variables having a common cause, rather than causing each other (Aldrich, 1995). A priori there were two potential common causes that we wished to control for: student ability and the effect of the second learning aid. We controlled for these potential confounds using partial correlations. Specifically, when measuring the partial correlation between participation with one learning aid and the final exam score, we controlled for performance on the two assignments and the number of days the other learning aid was accessed. Despite controlling for both these potential confounds we still found that the number of days Cram and PeerWise were accessed on was significantly correlated with the final exam mark.

A second potential limitation is that it is not clear to what extent our results can be generalized to other learning aids. To avoid overwhelming our students, we chose to trial only two learning aids. These learning aids were chosen to be representative of two broad classes of learning aid. For example, a number of learning aids employ electronic flashcards, so are very similar to Cram (e.g., www.quizlet.com, www.studyblue.com, www.flashcardmachine.com, www.scholastic.com etc.). As such, it is likely that our results from Cram would also apply to these learning aids. Similarly, a number of learning aids are similar to PeerWise in that they also rely on multiple choice quizzes. These include QPPA (Yu et al., 2002), AGQ (Chang et al., 2005), and Questionbank (Draaijer and Boter, 2005). For a review see Luxton-Reilly (2012). We expect that had we used one of these systems instead of PeerWise, we would have obtained similar results.

While we believe that our results are representative of these two broad classes of quiz-based learning aids, our results may not generalize to other quiz-based learning aids. For example, some learning aids use recall-based quizzes like Cram, but the quizzes are generated by students like PeerWise (Luxton-Reilly, 2012). StudySieve is an example of such a recall-based learning aid. While it was designed to be an improvement on PeerWise, it appears to be less effective. In particular participation with it is not correlated with improved exam scores (Luxton-Reilly et al., 2012). This is further evidence that recall-based learning aids may not be superior to recognition-based learning aids in an actual university setting.

Finally, we note that the two revision aids focused only on early stages of learning. According to the revised Bloom's taxonomy, there are six stages of learning: remembering, understanding, applying, analyzing, evaluating and creating (Anderson and Krathwohl, 2001). According to Robert Marzano's new taxonomy, which was also designed to overcome some of the shortcomings of the original Bloom's taxonomy, learning is achieved by the cognitive system in four stages: retrieval, comprehension, analysis and knowledge utilization (Marzano, 2001). According to either system, the two learning aids assist only the first form of learning: remembering (revised Bloom taxonomy) or retrieval (Marzano's new taxonomy). As such, they are likely to be helpful only for superficial learning.

CONCLUSION

In summary, we found that both computer-based learning aids were effective in that participation with either learning aid was correlated with higher performance in the final exam even when student ability and participation with the other learning aid was discounted. However, against our expectations, Cram did not outperform PeerWise. This is surprising since Cram is a recall-based learning aid whereas PeerWise is a recognition-based learning aid and recall-based learning aids have been consistently shown in the laboratory setting to result in more learning and better performance on the final quiz/test than recognition-based learning aids (Bjork and Whitten, 1974; Glover, 1989; Carpenter and DeLosh, 2006; McDaniel et al., 2013). Neither the format of the final exam, the degree of confidence engendered by the learning aids or the differing degrees to which the two learning aids engaged the students seems to be able to explain why the correlations between participation with Cram and final exam performance was not significantly greater than the correlation between participation with PeerWise and the final exam performance. Our results indicate that these laboratory findings do not necessarily apply to a university setting, possibly because in a realistic setting where more complicated material needs to be learned, recall-based learning aids provide lower quality feedback than that provided by recognition-based learning aids. Additionally, commercially available recall-based learning aids generally do not force the student to answer the questions before receiving feedback whereas recognition-based learning aids usually do. Thus, the later can enforce more engagement with the material than the former, so may result in more learning. Based on our findings, we believe that university instructors should not favor recall-based learning aids. As recognition-based learning aids are easier to administer, we would suggest that instructors should use them instead. We further note that with some recognition-based learning aids (e.g., PeerWise) it is possible to have students generate the questions in a collaborative fashion. This greatly

cuts down on the demands on the instructor/tutor responsible for administering the system. In addition, having students generate learning questions for each other is known to aid their learning (Slamecka and Graf, 1978). This strategy does not seem to work with recall-based learning aids (Luxton-Reilly et al., 2012), presumably because with recall-based learning aids it is harder to give high-quality feedback, so it is harder to construct suitable questions. It seems that students struggle to do this.

AUTHOR CONTRIBUTIONS

All authors helped plan the study. PH ran the study and analyzed the data. All authors helped write the article.

FUNDING

This work was supported by a Learning and Teaching Initiative Grant from the University of Melbourne.

REFERENCES

Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. *Stat. Sci.* 10, 364–376. doi: 10.1214/ss/1177009870

Anderson, L. W., and Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York, NY: Longman.

Bjork, R. A., and Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cogn. Psychol.* 6, 173–189. doi: 10.1016/0010-0285(74)90009-7

Carpenter, S. K., and DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: support for the elaborative retrieval explanation of the testing effect. *Mem. Cognit.* 34, 268–276. doi: 10.3758/BF03193405

Carpenter, S. K., Lund, T. J. S., Coffman, C. R., Armstrong, P. I., Lamm, M. H., and Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educ. Psychol. Rev.* 28, 353–375. doi: 10.1007/s10648-015-9311-9

Chang, S.B., Huang, H. M., Tung, K. J., and Chan, T. W. (2005). "AGQ: a model of student question generation supported by one-on-one educational computing. *CSCL '05*," in *Proceedings of the 2005 Conference on Computer Support for Collaborative Learning* (Taipei: International Society of the Learning Sciences), 28–32.

Denny, P., Hamer, J., Luxton-Reilly, A., and Purchase, H. (2008a). "PeerWise: students sharing their mulitiple choice questions," in *Proceedings of the Fourth International Workshop on Computing Education Research* (Sydney, NSW), 51–58.

Denny, P., Luxton-Reilly, A., and Hamer, J. (2008b). "The PeerWise system of student contributed assessment questions," in *Proceeding of the Tenth Conference on Australasian Computing Eduction* (Wollongong, NSW), 69–74.

Draaijer, S., and Boter, J. (2005). "Questionbank: computer supported self-questioning," in *9th CAA International Computer Assisted Assessment Conference* (Leicestershire, UK).

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., and Willingham, D. T. (2013). Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychol. Sci. Public Interest.* 14, 4–48. doi: 10.1177/1529100612453266

Efron, B., and Tibshirani, R. J. (1998). *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall/CRC.

Glover, J. A. (1989). The "testing" phenomenon: not gone but nearly forgotten. *J. Educ. Psychol.* 81, 392–399. doi: 10.1037/0022-0663.81.3.392

Hardy, J., Bates, S. P., Casey, M. M., Galloway, K. W., Galloway, R. K., Kay, A. E., et al. (2014). Student-generated content: enhancing learning through sharing multiple-choice questions. *Int. J. Sci. Educ.* 36, 2180–2194. doi: 10.1080/09500693.2014.916831

Kang, S. H. K., McDermott, K. B., and Roediger, H. L. I. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *Eur. J. Cogn. Psychol.* 19, 528–558. doi: 10.1080/09541440601056620

Karpicke, J. D., Blunt, J. R., Smith, M. A., and Karpicke, S. S. (2014). Retrieval-based learning: the need for guided retrieval in elementary school children. *J. Appl. Res. Mem. Cogn.* 3, 198–206. doi: 10.1016/j.jarmac.2014.07.008

Karpicke, J. D., and Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science* 319, 966–968. doi: 10.1126/science.1152408

Liaw, S.-S., Huang, H.- M., and, Chen, G.- D. (2007). Surveying instructor and learner attitutes towards e-learning. *Comput. Edu.* 49, 1066–1080. doi: 10.1016/j.compedu.2006.01.001

Lodge, J. M. and Horvath, J. C. (2017). "Science of learning and digital learning environments," in *From the Laboratory to the Classroom - Translating Science of Learning for Teachers Horvath*, eds J. C. Lodge, J. M. Lodge, and J. Hattie (Abingdon; Oxfordshire: Routledge), 122–135.

Luxton-Reilly, A. (2012). *The Design and Evaluation of StudySieve: A Tool that Supports Student-generated Free-Response Questions, Answers and Evaluations*. Ph.D, Auckland: The University of Auckland.

Luxton-Reilly, A., Bertinshaw, D., Denny, P., Plimmer, B., and Sheehan, R. (2012). "The impact of question generation activities on performance," in *Proceedings of the 16th Annual Joint Conference on Innovation and Technology in Computer Science Education* (New York, NY: Association for Computing Machinery).

Marzano, R. J. (2001). *Designing a New Taxonomy of Education Objectives. Experts in Assessment*. Thousand Oaks, CA: Corwin Press.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., and Morrisette, N. (2007). Testing the testing effect in the classroom. *Eur. J. Cogn. Psychol.* 19, 494–513. doi: 10.1080/09541440701326154

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., and Roediger, H. L. I. (2013). Quizzing in middle-school science: successful transfer performance on classroom exams. *Appl. Cogn. Psychol.* 27, 360–372. doi: 10.1002/acp.2914

McQueen, H. A., Shields, C., Finnegan, D. J., Higham, J., and Simmen, M. W. (2014). PeerWise provides significant academic benefits to biological science students across diverse learning tasks, but with minimal instructor intervention. *Biochem. Mol. Biol. Edu.* 42, 371–381. doi: 10.1002/bmb.20806

Nguyen, K., and McDaniel, M. A. (2015). Using quizzing to enhance student learning in the classroom: the good, the bad and the ugly. *Teach. Psychol.* 42, 87–92. doi: 10.1177/0098628314562685

Rawson, K. A., and Dunlosky, J. (2011). Optimising schedules of retrieval practice for durable and efficient learning: how much is enough? *J. Exp. Psychol. Gen.* 140, 283–302. doi: 10.1037/a0023956

Rhodes, J. (2013). *Using PeerWise to Knowledge Build and Consolidate Knowledge in Nursing Education*. Invercargill: Southern Institute of Technology Journal of Applied Research.

Robinson, R., Molenda, M., and Rezabek, L. (2016). *Facilitating Learning*. Bloomington, IN: Association for Educational Communications and Technology.

Roediger, H. L., and Butler, A. C. (2011). The crtical role of retrieval practice in long-term retention. *Trends Cogn. Sci. (Regul. Ed).* 15, 20–27. doi: 10.1016/j.tics.2010.09.003

Roediger, H. L., and Karpicke, J. D. (2006). Test-enhanced larning: taking memory tests improves long-term retention. *Psychol. Sci.* 17, 249–255. doi: 10.1111/j.1467-9280.2006.01693.x

Roediger, H. L. I., Putnam, A. L., and Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. *Psychol. Learn. Motivat.* 55, 1–36. doi: 10.1016/B978-0-12-387691-1.00001-6

Rohrer, D., Taylor, K., and Sholar, B. (2010). Tests enhance the transfer of learning. *J. Exp. Psychol. Learn. Mem. Cogn.* 36, 233–239. doi: 10.1037/a0017678

Slamecka, N. J., and Graf, P. (1978). The generation effect: delineation of a phenomenon. *J. Exp. Psychol. Human Learn. Mem.* 4, 592–604. doi: 10.1037/0278-7393.4.6.592

Trumbo, M. C., Leiting, K. A., McDaniel, M. A., and Hodge, G. K. (2016). Effects of reinforcement on test-enhanced learning in large, diverse, introductory college psychology course. *J. Exp. Psychol. Appl.* 22, 148–160. doi: 10.1037/xap0000082

Van Gog, T., and Sweller, J. (2015). Not new, but nearly forgotten: the testing effect decrease or even disappears as the complexity of learning materials increases. *Educ. Psychol. Rev.* 27, 247–264. doi: 10.1007/s10648-015-9310-x

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., and Liu, Y. (2014). The testing effect with authentic eduational materials: a cautionary note. *J. Appl. Res. Mem. Cogn.* 3, 214–221. doi: 10.1016/j.jarmac.2014.07.001

Yu, F. Y., Liu, Y. H., and Chan, T. H. (2002). "The efficacy of a web-based domain independent question-posing and peer assessment learning system," in *International Conference on Computers in Education* (Auckland), 641–642.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.