EUROPEAN JOURNAL OF COGNITIVE PSYCHOLOGY
2007, 19 (4/5), 514–527

Psychology Press
Taylor & Francis Group

Testing improves long-term retention in a simulated classroom setting

Andrew C. Butler and Henry L. Roediger, III
Washington University in St Louis, St Louis, MO, USA

The benefits of testing on long-term retention of lecture material were examined in a simulated classroom setting. Participants viewed a series of three lectures on consecutive days and engaged in a different type of postlecture activity on each day: studying a lecture summary, taking a multiple choice test, or taking a short answer test. Feedback (correct answers) was provided for half of the responses on the multiple choice and short answer tests. A final comprehensive short answer test was given 1 month later. Restudying or taking a multiple choice test soon after learning improved final recall relative to no activity, but taking an initial short answer test improved final recall the most. Feedback did not affect retention, probably due to the high level of performance on the initial tests. This finding is a powerful demonstration of how tests (especially recall tests) can improve retention of material after long retention intervals.

In most educational settings, tests are employed as a means to evaluate student learning for the purpose of assigning grades. The heavy emphasis on assessment often obscures another function of testing that is highly relevant to the goals of education: the promotion of learning. Considerable research in cognitive psychology has demonstrated that testing improves retention of the material tested, a phenomenon called the testing effect (Carrier & Pashler, 1992; McDaniel & Masson, 1985; Wheeler & Roediger, 1992; see Roediger & Karpicke, 2006a, for a review). To be sure, the idea of using tests as a learning tool in the classroom is not new (Gates, 1917; Jones, 1923–1924; Spitzer, 1939), and many researchers have made a case for the benefit of frequent testing in education (Bangert-Drowns, Kulik, & Kulik, 1991; Foos & Fisher, 1988; Glover, 1989; Leeming, 2002; Paige, 1966). However,

Correspondence should be addressed to Andrew C. Butler, Department of Psychology, Campus Box 1125, Washington University, 1 Brookings Drive, St Louis, MO 63139-4899, USA. E-mail: butler@wustl.edu

We thank Aurora Steinle for her help in creating the experimental materials and collecting data. This research was supported by a grant from the Institute of Education Sciences (No. R305H030339).

http://www.psypress.com/ecp DOI: 10.1080/09541440701326097

many of the laboratory studies that demonstrate the benefits of testing utilise basic materials, such as word lists, and retention intervals that usually are quite modest, such as a test at the end of a single experimental session or at most spanning a couple of days (e.g., Allen, Mahler, & Estes, 1969; Hogan & Kintsch, 1971; Thompson, Wenger, & Bartling, 1978). In the effort to apply the benefits of testing to educational practice, an important question remains: to what extent can findings from the laboratory be transferred to the classroom?

Jones (1923–1924) was the first researcher to investigate this question by conducting a series of experiments to study the retention of lecture material in the college classroom. Alarmed by the poor retention of lecture material he found in his first set of experiments (on average only two-thirds of the material was recalled on an immediate test and markedly less after a delay), he decided to assess whether previous findings about the benefits of recitation (Gates, 1917) could be applied to the college classroom. In perhaps his most impressive experiment, Jones investigated the effect of testing on later retention by giving students a brief completion test (e.g., fill-in-the-blank, short answer) immediately after a one hour class lecture and then retesting them after various delays (3 days to 8 weeks) to measure how much of the material they had forgotten. His control condition (for the purpose of comparison with the retest score) was a test of equivalent delay that covered material from the same lecture that had not been previously tested. The data, collected from 600 students across 27 lecture sessions, revealed a large benefit of testing: The amount of information retained after 8 weeks with a prior test was greater than that retained after just 3 days without a prior test. Overall, Jones concluded that testing is an effective method for improving the retention of lecture material and also indicated that tests should be given immediately to maximise their benefit (see also Spitzer, 1939).

The experiments conducted by Jones (1923–1924) are groundbreaking in that he used educationally relevant materials (class lectures) and long retention intervals (up to 8 weeks) to provide solid evidence that tests can be used as learning tools in the classroom. However, one problem with drawing firm conclusions from the results of the study is his failure to equate for total exposure time to the material for the two groups. That is, testing may simply have permitted students to selectively restudy the recalled material, so the benefit from testing could be due just to such restudying. In more recent research on the testing effect, a control group that restudies the material has been employed to equate for overall processing time in order to negate the hypothesis that testing is beneficial only because it involves additional exposure to the material (e.g., Roediger & Karpicke, 2006b). Interestingly, Jones did compare testing to additional study in a separate experiment using paired associates, but chose not to incorporate this design feature in his

experiment with class lecture materials (possibly due of the difficulty of producing an appropriate summary of an hour-long lecture).

Since Jones' (1923–1924) landmark study, a handful of subsequent experiments on the testing effect have used complex materials and longer retention intervals, but none has come close to combining the high degree of ecological validity and methodological rigor of his work (but see Metcalfe, Kornell, & Son, 2007 this issue). Many researchers have purposely incorporated educationally relevant materials in carefully controlled experiments with the goal of generalising to the classroom, a practice that dates from some of the first studies demonstrating the testing effect (e.g., Gates, 1917; Spitzer, 1939) to more recent efforts that have revived this tradition (e.g., Roediger & Karpicke, 2006b). Among the types of materials that have been used are foreign language paired associates (Carrier & Pashler, 1992), general knowledge questions (McDaniel & Fisher, 1991; Butler, Karpicke, & Roediger, 2007), and prose passages (Duchastel & Nungester, 1981; Foos & Fisher, 1988; Glover, 1989; LaPorte & Voss, 1975; Roediger & Karpicke, 2006b). Although the use of complex verbal materials in laboratory studies has strengthened the rationale for applying testing as a learning tool in education, even the most complex verbal materials (e.g., prose passages) are still relatively simple compared to the rich array of information encountered by students in the classroom.

Investigations of the testing effect that incorporate a retention interval of more than a week are rare and, to our knowledge, almost all of these studies have utilised naturalistic methodology to examine the extent to which information is retained over long periods of time. A prime example is the literature on the long-term retention of knowledge acquired in classrooms (e.g., Landauer & Ainsle, 1975; Semb, Ellis, & Araujo, 1993; for review see Semb & Ellis, 1994). Taken as a whole, these studies suggest that classroom testing benefits long-term retention of course material across a range of disciplines (e.g., medical education, physics, language instruction, etc.). However, instead of manipulating testing as an independent variable, these studies use testing to examine the retention of information over the period between a final course exam and a subsequent retention exam (often given as an afterthought) as a function of other variables, such as instructional technique and degree of original learning. In addition, these studies were conducted in real classrooms using established curriculum, a situation that introduces numerous uncontrolled factors (e.g., studying outside the classroom) and a lack of random assignment to groups (because ethical objections about placing students in a true control group). Another relevant example is research that investigates the maintenance of knowledge over retention intervals of many years. Bahrick and his colleagues have produced some of the best research on this topic showing the long-lasting benefits of testing (Bahrick, 1979; Bahrick, 1984; Bahrick & Hall, 1991). However, one

limitation of his studies is that he must rely on estimations of original learning in order to make feasible decade-long retention intervals. A cross-sectional design has also been used to study the long-term retention of knowledge learned in a cognitive psychology course (Conway, Cohen, & Stanhope, 1991).

Of the few studies that have manipulated testing as an independent variable to examine retention over longer intervals, almost all have used the relatively simple verbal materials described above (e.g., Nungester & Duchastel, 1982; Spitzer, 1939). One notable exception is a recent study by McDaniel, Anderson, Derbish, and Morrisette (2007, this issue) that investigated the benefits of testing over a semester using complex verbal materials. Students in a web-based course on "Brain and Behavior" were assigned 40 pages of reading per week and took either a short answer quiz, a multiple choice quiz, or read the facts that were used for the quiz conditions. Taking an initial short answer quiz led to superior performance on a subsequent multiple choice unit test relative to taking an initial multiple choice quiz or reading key facts.

The present experiment attempts to build upon the earlier work of Jones (1923–1924) by investigating the benefits of testing in a simulated college classroom setting. The study combined the experimental control of the laboratory with materials (art history lectures) like those found in a college classroom. We also used a long retention interval (1 month) to provide insight into a more realistic timescale over which students may retain classroom lecture information prior to a test. In addition to incorporating a "study" control group to equate presentation with the testing groups for total exposure to the materials, we investigated how different types of test (multiple choice and short answer) and the provision of feedback (correct answers given or not) would benefit retention of lecture material.

Participants watched a series of three lectures on consecutive days and engaged in a different learning activity after each lecture: taking a multiple choice test, taking a short answer test, or studying a lecture summary that contained points tested in other conditions. Each learning activity incorporated information from the lecture viewed that day only. Correct answer feedback (a presentation of the question and correct response) was given for half of the responses on the multiple choice and short answer tests.

One month later, participants returned for a comprehensive short answer exam that covered all three lectures. This final test included questions about information covered in the learning activities as well as information that had appeared in the lectures but that had not been re-presented during the any of the learning activities. This material from the lecture that was not presented again in any condition serves as a baseline against which to assess the effects of restudying or taking a multiple choice or short answer test.

METHOD

Participants and design

Twenty-seven Washington University undergraduates participated in the experiment (six other participants completed the initial sessions, but chose not to return for the final session and were therefore replaced by new participants). Course credit was given for the initial three learning sessions and a payment of $10 was given for the final test session. Participants were tested in groups of two to six people. The experiment employed a 2 (type of postlecture activity: multiple choice, short answer) × 3 (provision of test/feedback: no test, test without feedback, test with feedback) within-participants design. We also included an additional study control condition (a third type of postlecture activity) that could not be crossed with the provision of test/feedback factor. Thus, the overall design was unbalanced, but the experiment was fully counterbalanced and utilised a completely within-participants design. The type of postlecture activity factor and the additional study control were manipulated between lectures, whereas the provision of test/feedback factor was manipulated within lectures, but between items. That is, for each lecture in a testing condition, 10 items were not tested, 10 items were tested without feedback, and 10 items were tested with feedback.

Materials

Materials consisted of three videotaped lectures on art history from a series entitled *From Monet to Van Gogh: A history of impressionism* (The Teaching Company, 2000). The videos depicted a professor (Dr Richard Brettel) lecturing into the video camera (as if speaking to a classroom of students) interspersed with slides of relevant pieces of art and photographs. Each lecture covered the life and work of a single artist (Berthe Morisot, Auguste Renoir, Edgar Degas) and lasted 30 min.

For the purpose of the postlecture activities, 30 facts were selected from each lecture to create study and test materials. These facts covered many types of information (e.g., names, dates, events, etc.) and the timing of their presentation during the course of the lecture was evenly distributed over the 30 min. Lecture summary materials (for the study condition) were constructed by grouping the facts into paragraphs. Test materials were constructed by converting the facts into question/answer format. For example, a question from the Morisot lecture was "What aspect of Morisot's art could be used to date her paintings?" (Answer: *The fashions worn by the women*). For the purpose of multiple choice test, three plausible lures were developed for each question.

The experiment was counterbalanced in several ways. First, the 30 facts/questions for each lecture were divided into three sets of 10 items: Sets A, B, and C. To create the sets, the facts/questions were arranged by order of presentation in the lecture and randomly assigned to a set with the constraint that no set could receive more than one item from each consecutive group of three items. This method ensured that each set contained items that were evenly distributed over the course of the lecture. Second, three lecture presentation orders were created to counterbalance the sequence in which participants would view the lectures in the three initial learning sessions. The three orders were constructed such that overall each lecture would be presented equally often in each presentation position: (1) Renoir/Morisot/Degas, (2) Degas/Renoir/Morisot, (3) Morisot/Degas/Renoir. Third, three orders of the postlecture learning activities were created to counterbalance the sequence in which participants would engage in the different tasks. These orders were established such that across participants each activity occurred equally often after each session: (1) multiple choice/short answer/study, (2) study/multiple choice/short answer, (3) short answer/study/multiple choice. Finally, the counterbalancing orders for item set, lecture, and postlecture learning activity were factorially combined to create a total of 27 versions of the experiment. Each of the 27 participants was randomly assigned to one of these 27 versions.

Procedure

The experiment consisted of three initial learning sessions, which occurred on successive days, and a final test session, which took place about 1 month (28 days) after the final learning session. None of participants reported any prior experience with the material (e.g., an art history course on Impressionism).

Initial learning sessions. At the first session, participants were given a general overview of the experiment. Before watching the video, they were instructed to approach the lecture as they would a regular class and to take notes on blank paper that was provided. Although each participant took notes during all three initial sessions, the instruction to take notes was included to enhance the simulation of a classroom experience and therefore the notes were not subjected to any further analysis. When everyone was ready to begin, the video lecture was presented on a large screen at the front of the room by way of a mounted projector. After the lecture, the participants handed in their notes and moved to a computer to engage in the postlecture learning activity: studying a summary of the lecture, taking a multiple choice test, or taking a short answer test (depending on the task to

which they were assigned for that session). All the postlecture learning activities were presented individually on a PC using E-Prime software (Schneider, Eschman, & Zuccolotto, 2002) and the specific instructions for the assigned activities were explained at the start of the computer program. The postlecture portion of the session lasted approximately 10 min. Both the multiple choice and short answer tests were self-paced and the 20 questions were presented in a random order determined by the program (the other 10 questions associated with the lecture were not tested). Before the short answer test, participants were instructed to provide an answer to every question and told that any given answer should be no more than a sentence in length. Responses were entered using the keyboard. On both types of test, participants rated the confidence in their response after each question on a 4-point scale: 0 = guess, 1 = low confidence, 2 = medium confidence, or 3 = high confidence. After the confidence rating, participants saw either the correct answer feedback or a screen with "loading next question" for 6 s after each question (depending on the condition to which the item was assigned), so that total time spent on each question was roughly equated.

The study task consisted of reading a summary of the lecture that included all 30 facts from the lecture. Participants were instructed to read through the summary and pick up any information they had missed in the lecture. For the purpose of presentation, the summary was split up into three sections. Each section was displayed for 90 s (sufficient time to read through text once) before the program automatically cycled on to the next section. In total, the summary was presented twice (two complete cycles of the three sections) to keep participants engaged for the full duration of the postlecture activity. Thus, the time spent on each of the different postlecture activities was roughly equated with each task lasting approximately 10 min. The subsequent two learning sessions followed the same format: Participants watched a lecture (30 min) and then engaged in one of three postlecture learning activities (study, multiple choice test, short answer test). At the end of the third learning session, they were reminded about the final session and dismissed.

Final test session. Approximately 1 month after the third learning session, participants returned to take the comprehensive, self-paced, short answer test. The test consisted of 90 questions and covered all three lectures. As before, the test was given on a PC computer and the questions were presented in random order. Responses were entered using the keyboard. Instructions against guessing were given ("please answer only if you are reasonably sure you are correct") and thus omitting a response was identified as an option. After participants had finished the final test, they were debriefed and dismissed.

RESULTS

All results were significant at the .05 level of confidence unless otherwise noted. Pairwise comparisons were Bonferroni-corrected to the .05 level. In the analysis of repeated measures, a Geisser-Greenhouse correction was used for violations of the sphericity assumption (Geisser & Greenhouse, 1958).

Initial learning tests: Proportion correct

Overall, participants produced a high level of initial test performance: the proportion of correct responses on the multiple choice test ($M = 0.88$) was significantly higher than that of the short answer test ($M = 0.68$). However, this high level of performance was intended for two reasons: (1) to make sure that performance on the final test would be above floor, and (2) when using a test as a learning tool it is important that test-takers are able to retrieve a reasonable amount of the tested information, as Jones (1923–1924) and others have pointed out previously.

Final short answer test: Proportion correct

Figure 1 shows the proportion of correct recall for the final short answer test as a function of initial learning activity condition (data in the test conditions are collapsed across feedback conditions). The mean proportion correct for items in the no feedback and feedback conditions were almost identical for both types of prior test: multiple choice (no feedback = .36, feedback = .36)

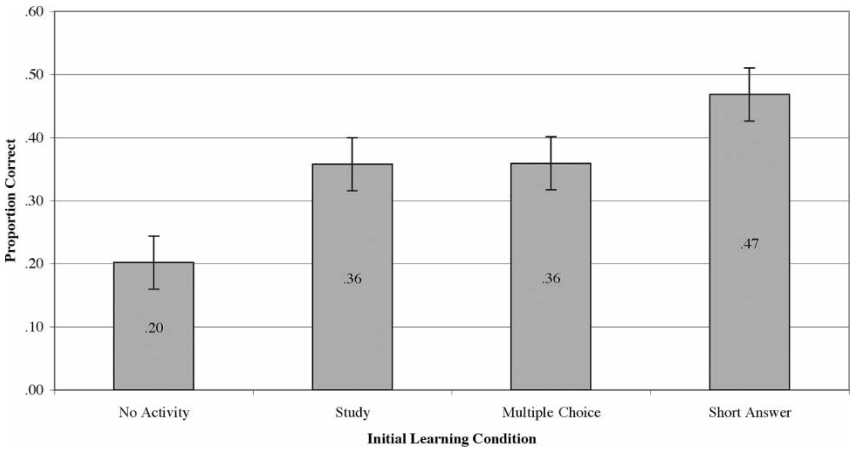

Figure 1. Mean proportion correct recall on the final short answer test as a function of initial postlecture learning condition (errors bars represent 95% confidence intervals).

and short answer (no feedback = .46, feedback = .47). This observation was confirmed by a 2 (initial test type: multiple choice, short answer) × 2 (provision of feedback: no feedback, feedback) repeated measures ANOVA in which there was no difference between provision of feedback conditions, $F(1, 26) = 0.01$, $MSE = 0.019$, $p = .95$. There was a significant main effect of initial test type, $F(1, 26) = 15.96$, $MSE = 0.020$, partial $\eta^2 = .38$, where taking a prior short answer test ($M = 0.47$) led to superior performance relative to a prior multiple choice test ($M = 0.36$). The interaction of initial test type and provision of feedback was not significant, $F(1, 26) = 0.09$, $MSE = 0.023$, $p = .76$. Thus, for the purpose of subsequent analysis, the data in the prior testing conditions were collapsed across the feedback conditions.

To examine the relative benefit of prior learning activity, a one-way repeated measures ANOVA was conducted with type of initial learning task (no test, study, multiple choice, short answer) as the factor and proportion correct as the dependent variable. This test revealed a significant difference among the four initial learning task conditions, $F(3, 78) = 27.07$, $MSE = 0.012$, partial $\eta^2 = .51$. Pairwise comparisons indicated that a higher proportion of items in the short answer condition were recalled than in either the multiple choice condition, $t(26) = 3.99$, $SEM = 0.027$, or the study condition, $t(26) = 3.17$, $SEM = 0.035$. There was no difference between the multiple choice and study conditions, $t(26) = .04$, $SEM = 0.031$, $p = .97$, but the study condition (and the other conditions) led to a higher proportion of correct responses than the no test condition, $t(26) = 5.10$, $SEM = 0.031$.

Final short answer test: Performance as a function of initial confidence

Performance for the two initial test conditions (multiple choice and short answer) was broken down as a function of initial confidence estimates. Due to the high level of initial test performance, confidence estimates were skewed towards the "high confidence" end of the scale. Overall, higher levels of initial confidence led to a higher proportion correct on the final short answer test. However, there were no systematic differences between the two feedback conditions at any of confidence levels.

GENERAL DISCUSSION

This experiment examined how different types of postlecture activity affected retention of lecture material over a realistic (1 month) retention interval as measured by a final short answer test. We found that taking a prior short answer test produced significantly better retention of the material than both studying a lecture summary or taking a multiple choice test.

Although there was no difference in the amount of material retained in the study and multiple choice conditions, all three conditions in which participants engaged in a postlecture activity resulted in superior performance relative to the no activity condition. Surprisingly, the provision of feedback after responses did not improve retention of the material in either of the test activity conditions (multiple choice and short answer). We now turn to discussing each of these results.

The primary finding was that taking a short answer test produced superior retention of lecture material after 1 month relative to studying a lecture summary, a control condition in which participants were essentially shown twice all the critical facts that would later be tested. This result provides compelling evidence that testing can improve the retention of classroom lecture material by way of a postlecture test procedure that can be easily implemented in the classroom. Many studies have found that taking a test leads to greater retention of the material relative to a restudy condition, especially when the test involves response production (e.g., Duchastel & Nungester, 1981; Hogan & Kintch, 1971; McDaniel et al., 2007 this issue; Roediger & Karpicke, 2006b; Thompson et al., 1978).

The short answer test condition also produced superior retention relative to the multiple choice condition. This result fits well with previous laboratory research using basic materials that shows that taking an initial recall test confers larger benefits on subsequent test than taking an initial recognition test (Cooper & Monk, 1976; McDaniel & Masson, 1985; Wenger, Thompson, & Bartling, 1980). This result is also consistent with research using educationally relevant materials in which an initial short answer test produces superior performance on a subsequent test relative to an initial multiple choice test (Duchastel, 1981; Kang, McDermott, & Roediger, 2007 this issue; McDaniel et al., 2007 this issue), a result that generally occurs regardless of whether the final test is in short answer or multiple choice format (e.g., Foos & Fisher, 1988; Glover, 1989; Kang et al., 2007 this issue). Some studies have found an overall superiority of initial multiple choice test, but this may be due to very low performance on the initial short answer test (e.g., Kang et al., 2007 this issue, Exp. 1). Theoretically, one explanation for these results is the idea that greater depth or difficulty in retrieval leads to better retention of the information tested (Bjork, 1975; McDaniel & Masson, 1985). Presumably, short answer tests (which require the production of a response) involve a greater degree of retrieval difficulty than multiple choice tests (which require the selecting the correct response from a number of alternatives). In addition, the results could be explained within a transfer-appropriate processing framework (Morris, Bransford, & Franks, 1977). Taking an initial short answer test would be expected to promote better transfer relative to an initial multiple choice test when the final test is a short answer test. It is important to note

that these two theoretical explanations are not mutually exclusive, and both likely play a role in producing the present results.

The finding that the short answer test condition produced better retention than the no activity control condition is notable in that it replicates the findings of Jones (1923–1924) as well as other previous studies (Duchastel, 1980; Glover, 1989; Wheeler & Roediger, 1992). Interestingly, performance on the final short answer test was equivalent for the multiple choice and study groups. A possible explanation is that the lecture summary provided in the study condition gave a distinct advantage by exposing participants twice to all the critical facts that they would later be tested on. With respect to educational practice, this is a rather artificial study task because educators would never give students the answers to the test ahead of time. A more realistic control condition, and one we recommend for future studies of this type, would be to permit students to review the notes they took during the lecture.

One puzzling finding is that feedback did not improve retention of the material. In many experiments, feedback has a profound effect on retention (e.g., McDaniel & Fisher, 1991) because it helps test-takers to correct errors (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Pashler, Cepeda, Wixted, & Rohrer, 2005) and to confirm correct responses (Butler et al., 2007). This null effect is likely due in part to the high level of performance on the initial tests, especially for multiple choice: Feedback was not as useful because few errors were made (see Kang et al., 2007 this issue). However, this reasoning cannot fully explain why feedback did not have an effect on the initial short answer test as participants got almost a third of the responses incorrect ($M = 0.32$). Other factors that may have led to ineffectiveness of feedback were the amount of time participants were given to process the feedback and the fact that it occurred immediately after subjects responded. The information tested by any given question was quite complicated (e.g., an answer often consisted of a long phrase or sentence). Feedback was presented for only 6 s and this amount of time may not have been sufficient to allow participants to fully process the information. The timing of feedback may be critical, because evidence exists suggesting that the ratio between the interstudy interval and retention interval maximises retention (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). If feedback is conceptualised as an additional study opportunity (i.e., in addition to the initial exposure to the material), this research would suggest that should be presented after a delay in order to produce spaced presentations and optimal retention. Of course, an alternative hypothesis is that giving immediate feedback simply does not affect retention over long periods of time, but this generalisation seems unlikely because the type of feedback used (a representation of the question and the correct answer) almost always increases learning from tests (see Roediger & Karpicke, 2006a). On a related note,

feedback may be very important to reducing the negative effects that arise from exposing test-takers to misinformation in the form of multiple choice lures (Roediger & Marsh, 2005). However, in our experiment, very few lure items from the multiple choice test were produced as answers on the final short answer test ($M = 0.04$), and the proportion of lures produced on the final test in the initial multiple choice condition did not differ from the baseline rate of spontaneously producing these responses in the other conditions.

We believe the present findings have direct implications for educational practice. Our experiment combined ecologically valid presentation materials (actual lectures) and realistic retention intervals (1 month). This combination makes our study one of the most powerful demonstrations to date of how the mnemonic benefits of testing can be applied to enhance classroom learning. The benefit of taking a brief quiz (either short answer or multiple choice) is especially striking when compared with the no activity condition, which is perhaps more indicative of common practice in the classroom than the restudy condition. In addition to boosting retention, frequent testing can help to lower students' test anxiety and increase the regularity of studying (Leeming, 2002). Although it did not have an effect in the present study, feedback should also be provided to ensure students are learning from the test, especially in the event of poor test performance. To minimise the time taken away from the primary classroom activities, feedback could be accomplished by requiring students to self-correct their tests after the class period. We encourage educators to incorporate testing into their daily classroom routine: The amount of class time sacrificed for a quiz is small compared to gain in retention of material.

REFERENCES

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior*, *8*, 463–470.

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General*, *108*, 296–308.

Bahrick, H. P. (1984). Semantic memory content in permastore: 50 years of memory for Spanish learned in school. *Journal of Experimental Psychology: General*, *113*, 1–29.

Bahrick, H. P., & Hall, L. K. (1991). Lifetime maintenance of high school mathematics content. *Journal of Experimental Psychology: General*, *120*, 20–33.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research*, *85*, 89–99.

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*, 213–238.

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition* (pp. 123–144). New York: Wiley.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. III., (2007). A matter of confidence: Correct responses benefit from feedback. *Manuscript submitted for publication*.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory and Cognition*, *20*, 633–642.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*, 354–380.

Conway, M. A., Cohen, G., & Stanhope, N. (1991). On the very long-term retention of knowledge acquired through formal education: Twelve years of cognitive psychology. *Journal of Experimental Psychology: General*, *120*, 395–409.

Cooper, A. J. R., & Monk, A. (1976). Learning for recall and learning for recognition. In J. Brown (Ed.), *Recall and recognition* (pp. 115–140). London: Wiley.

Duchastel, P. C. (1980). Extension of testing effects on the retention of prose. *Psychological Reports*, *47*, 1062.

Duchastel, P. C. (1981). Retention of prose following testing with different types of test. *Contemporary Educational Psychology*, *6*, 217–226.

Duchastel, P. C., & Nungester, R. J. (1981). Long-term retention of prose following testing. *Psychological Reports*, *49*, 470.

Foos, P. W., & Fisher, R. P. (1988). Using tests as learning opportunities. *Journal of Educational Psychology*, *80*, 179–183.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, *6*, No. 40.

Geisser, S., & Greenhouse, S. W. (1958). An extension of Box's results on the use of F distribution in multivariate analysis. *Annals of Mathematical Statistics*, *29*, 885–891.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*, 392–399.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, *10*, 562–567.

Jones, H. E. (1923). The effects of examination on the performance of learning. *Archives of Psychology*, *10*, 1–70.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modulate the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*, 528–558.

Landauer, T. K., & Ainslie, K. I. (1975). Exams and use as preservatives of course-acquired knowledge. *Journal of Educational Research*, *69*, 99–104.

LaPorte, R. E., & Voss, J. F. (1975). Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology*, *67*, 259–266.

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology*, *29*, 210–212.

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*, 494–513.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, *16*, 192–201.

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *11*, 371–385.

Metcalfe, J., Kornell, N., & Son, L. K. (2007). A cognitive-science based programme to enhance study efficacy in a high- and low-risk setting. *European Journal of Cognitive Psychology*, *19*, 743–768.

Morris, P. E., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer-appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533.

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology*, *74*, 18–22.

Paige, D. D. (1966). Learning while testing. *Journal of Educational Research*, *59*(6), 276–277.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*, 3–8.

Roediger, H. L., III, & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.

Roediger, H. L., III, & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255.

Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequence of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*, 1155–1159.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime reference guide*. Pittsburgh, PA: Psychology Software Tools, Inc.

Semb, G. B., & Ellis, J. A. (1994). Knowledge taught in school: What is remembered? *Review of Educational Research*, *64*, 253–286.

Semb, G. B., Ellis, J. A., & Araujo, J. (1993). Long-term memory for knowledge learned in school. *Journal of Educational Psychology*, *85*, 305–316.

Spitzer, H. J. (1939). Studies in retention. *Journal of Educational Psychology*, *30*, 641–656.

The Teaching Company (Producer). (2000). *From Monet to Van Gogh: A history of impressionism* [Motion picture]. (Available from The Teaching Company, Chantilly, VA)

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 210–221.

Wenger, S. K., Thompson, C. P., & Bartling, C. A. (1980). Recall facilitates subsequent recognition. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 545–559.

Wheeler, M. A., & Roediger, H. L., III. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science*, *3*, 240–245.