SPACED TESTING AND MEMORY RETENTION FOR TEXTBOOK READINGS

Assistant Professor Dr. Kyle J. Susa, Ph.D. & Steven Dessenberger
Department of Psychology, Kegley Institute of Ethics Faculty Fellow, California State University, Bakersfield
ksusa@csub.edu

Abstract: Known as the *testing effect*, decades of educational research in psychology have demonstrated that testing memory enhances subsequent memory retention. This effect is bolstered when testing is repeated over time at spaced intervals. The current study examined whether test repetition and various intervals of spacing could be implemented into a college course to enhance students' memory retention for textbook readings. Students completed weekly online tests of chapter readings over the course of the semester, whereby critical test questions from early in the semester were repeated 3 times (or once in the control condition) using either equal-spaced or expanded-spaced repetition intervals. Results indicate that students' memory for information read is best retained for questions administered using expanded-spaced intervals.

INTRODUCTION

Known as test-enhanced learning, empirical research has indicated that testing has a remarkable and robust impact on long-term memory retention. In fact, hundreds of studies in over a century of research have continually demonstrated that testing, compared to restudying or not testing, enhances memory retention and transfer of learning (Adesope, Trevisan, & Sundararajan, 2017; Rowland, 2014). Nonetheless, testing often has a negative connotation in public policy, and the benefits of testing as an evidence-based pedagogical practice are relatively unknown by teachers and students alike (Karpicke, Butler, & Roediger, 2009). Consequently, although professors are constantly striving to improve their courses, little class time is often devoted to what we know about the amount of effort and repetition necessary to remember information in the long term. This is most likely exacerbated in a field as broad as psychology, where course curriculum often encompasses much breadth. Furthermore, there are few, if any, formal graduation standards that require undergraduate students to remember what they have learned in any given course.

Despite a lack of attention on formal memory retention requirements in higher education, it could be argued that effective teaching only occurs when students remember what they are taught. Known as *desirable difficulties*, a long-standing principle in cognitive psychology suggests that memory for information learned is best retained through the creation of "difficult" conditions under which memory is retrieved (Bjork 1994; Pyc & Rawson, 2009). That is, creating challenging opportunities for students to retrieve class material (through testing), after they have initially encoded it (rather than simply rereading or highlighting notes), is essential for long-term retention. This *testing effect* is enhanced through *repeated testing* and *spaced learning* - the idea that testing (or studying) should be repeated more than once and spaced over a period of time (Karpicke & Roediger, 2007).

Basic and applied research on repeated testing of memory has routinely demonstrated strong effects on long-term memory retention (see Roediger & Butler, 2011 for a review). In one classic study, Wheeler and Roediger (1992) had participants remember pictures and tested them either one or three times (or not all in the control condition). Results indicated that participants who took three initial tests recalled more correct responses than participants in the other conditions. This effect has also been found in prose passage reading (Roediger & Karpicke, 2006), associative word learning (McDermott, 2006), and other domains of memory (Karpicke & Roediger, 2007).

More recently, much of repeated-testing (and repeated-studying) research has focused on the optimal level of spacing between test repetitions. In large part, this research has examined two distinct spaced-learning intervals. The first is an *equal* or *fixed* interval, where the time between repeated test sessions is always the same. For example, if the time between the first and second repetitions of the test is one week, the time between the second and third repetitions will also be one week. The second type of interval is often referred to as *expanded* or *incremental*, where the time between repeated test sessions gradually increases. In this instance, the time between the first and second repetitions of a test may be one week, whereas the time between second and third repetitions is three weeks, and the time between the third and fourth repetitions is five weeks.

In line with a desirable difficulty theoretical framework, many researchers believe that expanded intervals are superior to equal intervals because increased time (i.e., difficulty) requires the learner to use more effortful processing in retrieving the information, which subsequently leads to greater retention (Landauer & Bjork, 1978).

Nonetheless, this belief has mixed empirical support within the spaced-learning literature. Some studies demonstrate an equal spacing superiority effect (Karpicke & Roediger, 2007), whereas others suggest there are no discernable differences between spaced-learning intervals (Karpicke & Bauernschmidt, 2011; Storm, Bjork, & Storm, 2010).

Although researched less often, some studies have examined the effects of testing within classroom environments. This research suggests that effects found in the laboratory generally hold within the classroom (McDaniel, Roediger, & McDermott, 2007; McDaniel, Wildman, & Anderson 2012). For example, Pennebaker, Gosling, and Ferrell (2013) administered daily online quizzes in introductory courses and compared students' scores to those with more traditional testing procedures. They found that daily online quizzing resulted in better overall student performance in the course as well as in other courses that semester and in subsequent semesters. Atabek, Balkan, and Cetinkaya (2014) also found in a pretest–posttest design that having intervening multiple choice and matching tests in a chemistry course allowed students to retain information better than a control group. No studies, to our knowledge, have examined the effects of spaced repetitions of testing within the context of a college course, and few studies have examined the effects of repeated testing within a course on memory-retention intervals lasting longer than one week (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; McDaniel, Anderson, Derbish, & Morrisette, 2007).

The purpose of the present study was to examine the benefits of various intervals of spaced testing within a 16-week semester Psychology Research Methods course while measuring the retention of information learned three to five weeks after the final test repetition. Specifically, the study focused on the application of spaced-interval testing as a pedagogical strategy to enhance memory retention for information learned during weekly readings of the textbook. Multiple-choice reading questions from the first three weekly readings of the semester were repeated at one of three spaced intervals over nine weeks of the semester, before students took a final memory retention test during the final exam. Given the lack of research on spaced repetitions of testing in the context of an actual college course and the mixed findings in laboratory research (particularly with long retention intervals), our study was exploratory in nature. The study was purposefully designed to naturally flow within the structure of the course in order to presumably enhance students' retention of information read.

METHODOLOGY

Participants

The study was administered within normal pedagogical practices of the course. Participants included 49 undergraduate students at a state university in the United States of America. However, data were only analyzed from the 23 students who completed all of the weekly tests. Students consented to the release of their test data.

Materials

Weekly online tests were administered consisting of questions from the textbook, *Research Methods in Psychology* (Morling, 2015), via the learning management platform *Blackboard Learn*. The weekly tests consisted of 12–16 multiple-choice questions (1 correct answer, and 3 lures) from the relevant chapter readings for the week and critical questions. Test questions covered key terms and concepts of the course material. For example, "Some theories are better than others. Which of the following is NOT considered a feature of a good theory: a) the theory makes sense intuitively; b) the theory is supported by the data; c) the theory is parsimonious; d) the theory is falsifiable?"

The critical questions were questions that originated from the chapter readings during Weeks 2, 3, and 4 and were either repeated three times (at either equal or expanded intervals) or repeated once (in the control condition) over the course of the semester. In total, there were 36 critical questions, 12 that originated from each of the Weeks, 2, 3, and 4. Out of the 12 critical questions that originated from Weeks 2, 3, and 4, four questions were repeated using an equal-spaced schedule, four were repeated using an expanded-spaced schedule, and four were repeated once in the control condition. All of the critical questions were determined to be of approximate difficulty to each other, and it was randomly determined which questions were used in each of the spaced-testing intervals. The question order and the order of the answers (and lures) was randomized for every repetition of each question.

The final retention test was administered in paper and pencil format and was incorporated into the final exam of the course. The final exam consisted of 86 multiple-choice questions, 50 questions from lecture material and the 36 critical questions from the textbook readings.

Design

Within-subjects, the study examined the effect of three different spaced-testing intervals (equal, expanded, control) on memory retention of learned information from textbook readings. Each of the spaced-testing intervals had their last repetition during Weeks 11, 12, and 13 (depending on whether the question originated during Week 2, 3, or 4). Memory retention was assessed during the final repetition of the critical questions (during Weeks 11, 12, and 13 of the course), and 3-5 weeks later (depending on when the last repetition occurred) on the final exam for the course (Week 16). Memory retention was analyzed in three ways: 1) the percentage correct for each of the three spaced-testing intervals on the last repetition; 2) the percentage correct for each of the three spaced-testing intervals on the final exam; and 3) the change in percentage correct between the last repetition and final exam, for each of the three spaced-testing intervals.

Procedure

Students were instructed at the beginning of the semester that they would be taking required weekly tests on the chapter readings. They were told some of the questions on the tests would be from the weekly reading and other questions would be repeated questions from previous tests. Upon answering the test questions from the first three weeks of the semester (starting on Week 2) students were provided feedback, including whether they got the answer right and what the correct response was supposed to be when they selected the wrong answer. Students were told to use the feedback to improve their performance on these items if they were to see them again on future tests. Feedback was not included after the first three weeks to control for mere exposure effects. The tests were assigned one week prior to their due date. Once the students began the test they had to complete it but were allowed as much time as needed.

The questions originating from Weeks 2, 3, and 4 resulted in the critical questions that were repeated at one of the three spaced-testing intervals over the course of the semester. Namely, each of the tests given during the first three weeks consisted of 12 unique multiple-choice questions that were not directly covered during the course lecture. The questions randomly assigned to the equal-spaced condition reappeared every three weeks for a 3–3–3 repetition schedule; for example, they initially appeared in Week 2 and then appeared again in conjunction with the regular chapter test questions in Weeks 5, 8, and 11. The questions randomly assigned to the expanded-spaced condition reappeared one week later, followed by another presentation three weeks later and again five weeks after that for a 1–3–5 repetition schedule. The randomly assigned control condition questions only received one repetition retrieval. Figure 1 depicts the difference between the three spaced-testing intervals, over the course of the semester.

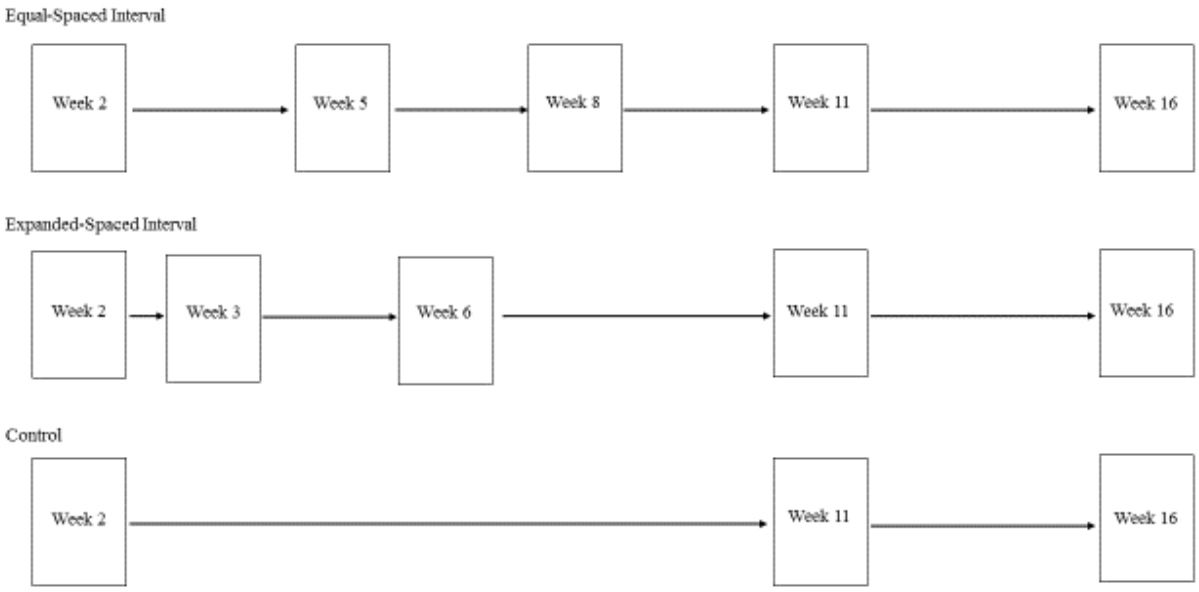

Figure 1. Repetition schedule of three spaced-testing intervals over the course of the semester.

Because the critical questions originated in either Week 2, 3, or 4, the repetition of questions was staggered throughout the weeks of the semester. The full staggered schedule of question repetitions is presented in Table 1. For example, equal-spaced questions that were initially retrieved during Week 3 were repeated at Weeks 6, 9, and 12. Equal-spaced questions that were initially retrieved during Week 4 were repeated at Weeks 7, 10, and 13. Further, expanded-spaced questions retrieved during Week 3 were repeated at Weeks 4, 7, and 12, while expanded-spaced questions retrieved during Week 4 were repeated at Weeks 5, 8, and 13. The spacing schedules were created so that both the equal and expanded conditions would have the third repetition of their questions occur in Weeks 11, 12, and 13 (depending on whether the questions originated from Week 2, 3, or 4). The control condition questions were also assessed during these weeks.

Table 1: *Staggered Schedule of Question Repetitions*

	Week of the Semester															
Critical Questions	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Originating from Week 2																
Equal		4			4R1			4R2			4R3					4F
Expanded		4	4R1			4R2					4R3					4F
Control		4									4R1					4F
Originating from Week 3																
Equal			4			4R1			4R2			4R3				4F
Expanded			4	4R1			4R2					4R3				4F
Control			4									4R1				4F
Originating from Week 4																
Equal				4			4R1			4R2			4R3			4F
Expanded				4	4R1			4R2					4R3			4F
Control				4									4R1			4F
Total Critical Questions	0	12	16	16	8	8	8	8	4	4	16	16	16	0	0	36

Note. R1 = first repetition; R2 = second repetition; R3 = third repetition, F = final retrieval

Prior to the final exam date, students were told that some of the questions would come from their chapter readings. The final exam consisted of 50 multiple-choice questions covering class lecture material and the 36 critical questions from the weekly reading tests. The critical questions were the last questions on the exam and were presented in random order. Students had ample time to complete the entire exam. Students were not allowed to use their textbook or notes on the final exam. After the exam, students were debriefed on the nature of the study and were offered the option of providing consent to release their test data for the purpose of the study.

RESULTS

Only students who completed all of the weekly tests were included the analyses (N = 23) so that all of the data properly reflected those that were administered the manipulations of interest. Results of the analyses are displayed in Figure 2.

Results of the last repetition test indicated there was no difference between any of the three conditions on students' memory retention during Weeks 11, 12, and 13, $F(2, 44) = .373$, $p = .691$, $\square_p^{\square} = .017$. Although it is somewhat surprising that the control condition (M = .92, SD = .12) performed as well as the equal (M = .90, SD = .16) and the expanded (M = .92, SD = .11) conditions, the results should be interpreted in context, as each condition had a different retention interval. Nonetheless, this comparison was important to understand the base-rate in performance between conditions before the final memory-retention test. It also allowed us to standardize the retention interval between the spaced-interval conditions, on the final memory-retention test.

The final memory-retention test occurred during the final exam period of the course (Week 16). Therefore, each of the three conditions had one-third of the questions with a three-week retention interval, one-third with a four-week

retention interval, and one-third with a five-week retention interval. For the purposes of statistical reliability, we collapsed the scores across the retention interval for each condition so that final performance was measured across 12 questions in each condition.

Results indicated there was a significant difference between spaced-testing intervals on memory retention, $F(2, 44) = 4.934$, $p = .012$, $\square_p^{\square} = .183$. Simple effect comparisons indicated that questions administered with an expanded interval ($M = .89$, $SD = .13$) were more accurately remembered than questions administered under the equal interval ($M = .82$, $SD = .20$), $t(22) = 3.029$, $p = .006$, $d = .42$, or in the control condition ($M = .85$, $SD = .15$), $t(22) = 2.614$, $p = .016$, $d = .28$. There was no difference in accuracy between the equal interval and the control condition, $t(22) = .935 =$, $p = .360$, $d = .17$.

To further examine the impact of the spaced intervals, the change in performance accuracy between conditions from the last repetition test to the final memory retention test was also assessed. Here, results indicated that students' scores for the equal interval, $t(22) = 2.420$, $p = .024$, $d = .44$, and control condition, $t(22) = 3.034$, $p = .006$, $d = .52$, resulted in a statistically significant drop in performance. However, students were able to retain the information they read when expanded-spaced testing was administered, $t(22) = 1.262$, $p = .220$, $d = .25$.

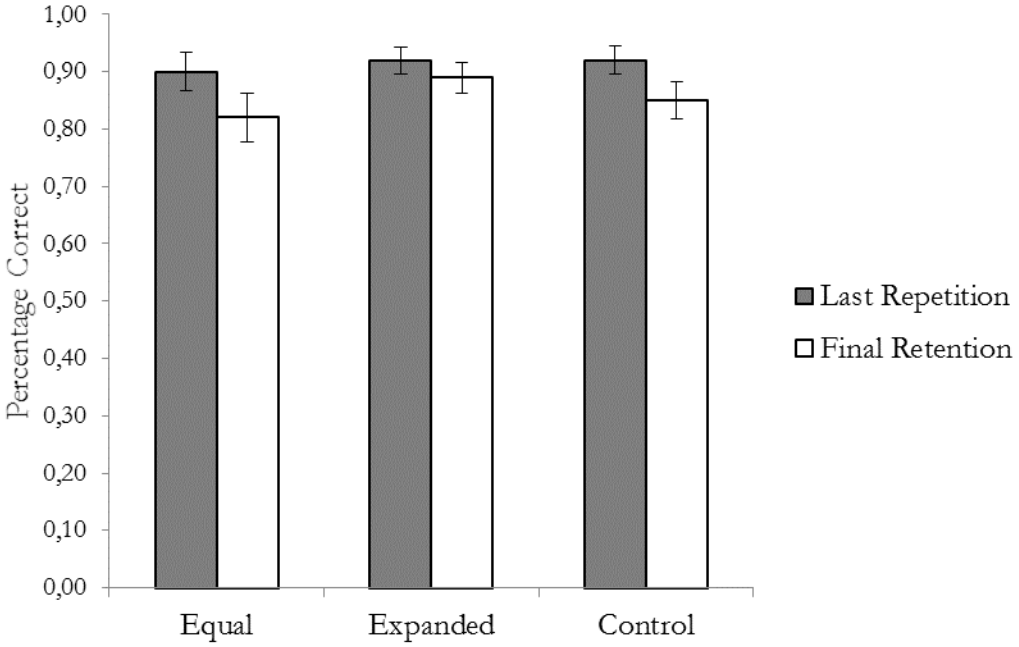

Figure 2. Memory retention as a function of time of retrieval and spaced-testing interval. Error bars represent standard errors.

These results suggest that within this paradigm, expanded-spaced testing is superior to both equal-spacing and control conditions in two important measures. First, when controlling for the retention interval, the overall accuracy rate is greater for expanded-spaced testing with a rather moderate effect size difference between both equal-spaced and control conditions. Second, a substantial drop in performance from the last repetition to the final exam was evident for both the equal-spaced and control questions, but not nearly as strong (nor statistically significant) in the expanded-spaced condition.

DISCUSSION

The present study used a novel semester-long experimental paradigm to investigate the effects of various repetition intervals of spaced testing on memory retention for textbook readings. This study contributes to both theoretical and pedagogical perspectives on how test-enhanced learning can be applied within a semester-long college course. Namely, findings from our study support a desirable-difficulties approach to testing, which suggests that effortful processing that arises from expanded-spaced retrieval is most beneficial for memory retention.

Two novel attributes of our study are the longitudinal nature with which the spaced intervals occurred and the length of the memory-retention interval. Although most testing-effect paradigms manipulate spacing and retention intervals over minutes or a few days, the spacing and retention intervals in the present study are conducive to a college semester and add to the debate on the optimal intervals of spacing when examining students' memory with a relatively long retention interval. It bears noting that although the results support other findings within the literature (e.g., Landauer & Bjork, 1978), they also contradict studies that show a superior effect for equal-spaced intervals (Karpicke & Roediger, 2007). Karpicke and Roediger (2007) argue that *retrieval lag* (the time between the initial encoding and first retrieval test) is an important factor to consider with spaced intervals. Compared to expanded intervals, equal-spaced intervals generally have greater retrieval lags, which can also lead to desirable difficulties. One could speculate that the relatively long (three-week) delay between the initial retrieval tests and the first repetition may have created a level of difficulty in the equal-spaced condition that was simply too great to be effective. It is important for future research to examine the interplay between retrieval lag, repetition intervals, and retention intervals in the context of a semester course. For example, what effect do different spaced-testing intervals have on memory retention when the delay between the initial test and the first repetition is held constant?

Of primary importance, the present study sought to advance evidenced-based pedagogical practices for enhancing student memory for textbook readings. The study was designed to fit the course, to require minimal extra effort on behalf of the instructor to implement, to not require any class time, and to make students' experience in the course consistent with regular expectations and requirements. Given this emphasis on practicality, the low-stakes nature of the reading tests was maintained, to give the students flexibility to take the online tests at their own pace, and to not force them to answer an abundance of questions during any given week. Two limitations of the study are important for future research. First, the test repetitions were not timed and students were not prohibited from taking the tests in an open-book format. One could speculate this may have impacted the strength of effortful retrieval in each of the conditions and may have enhanced the accuracy of the repetition test scores. Perhaps, a timed test would have led to lower scores overall, yet more pronounced differences in final memory retention between the conditions. In our study the open-book option did not systematically influence performance scores between conditions, especially on the final closed-book test. Further, Agarwal, Karpicke, Kang, Roediger, and McDermott (2008) demonstrated that the testing effect on final memory retention is prevalent when using either open or closed book initial retrieval tests.

A second limitation worthy of future research would be to include a true no repetition control group of questions. The condition was purposefully not included in our study to maintain consistency with typical course expectations. Further, the control that was included allowed for maintaining equal retention intervals between conditions on the final exam. Nonetheless, since the standard in most classes is to not repeat test questions after they have been tested once, it is an important empirical question to ask how either of the two spaced-testing intervals differs from a true no repetition control. Simply, the present study is a foundation for future research to explore the boundary conditions of this effect while considering the needs for practical implementation (e.g., number of test items, open- vs. closed-book tests, timed tests, depth of retrieval, presence of feedback, etc.) as well as how these factors might moderate the effects of space-testing intervals within this paradigm.

The present study provides encouraging data to suggest that an evidence-based approach to teaching can be easily implemented in a course curriculum to enhance students' memory retention for information they read. The approach is grounded in basic theory and methodologies of a desirable-difficulties perspective of memory retrieval, showing that expanded spaced-testing intervals of test items is best for retaining memory. It will be important for future researchers to advance our understanding of these effects, especially within the conditions of practical implementation. At present, this study provides a positive step forward - a step toward teaching for the sake of remembering.

REFERENCES

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Eductional Research, 87*(3), 659–701. doi:10.3102/0034654316689306

Agarwal, P. K., Karpicke, J. D., Kang, S. H., Roediger, H. L., III, & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22*, 861–876. doi:10.1002/acp.1391

Atabek Yigit, E., Balkan Kiyici, F. Y., & Cetinkaya, G. (2014). Evaluating the testing effect in the classroom: An effective way to retrieve learned information. *Eurasian Journal of Educational Research, 54*, 99–16.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

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*(3), 354–380. doi:10.1037/0033-2909.132.3.354

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(5), 1250–1257. doi:10.1037/a0023436

Karpicke, J. D., Butler, A., & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17*(4), 471–479. doi:10.1080/09658210802647009

Karpicke, J. D., & Roediger, H. L., III. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(4), 704–719. doi:10.1037/0278-7393.33.4.704

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. Gruneberg, P. Morris, & R. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). London, United Kingdom: Academic Press.

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. doi:10.1080/09541440701326154

McDaniel, M. A., Roediger, H. L., III, & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*(2), 200–206. doi:10.3758/BF03194052

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition, 1*(1), 18–26. doi:10.1016/j.jarmac.2011.10.001

McDermott, K. B. (2006). Paradoxical effects of testing: Repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition, 34*(2), 261–267. doi:10.3758/BF03193404

Morling, B. (2015). *Research methods in psychology*. New York, NY: W. W. Norton and Company, Inc.

Pennebaker, J. W., Gosling, S. D., & Ferrell, J. D. (2013). Daily online testing in large classes: Boosting college performance while reducing achievement gaps. *PLoS ONE, 8*(11) doi:10.1371/journal.pone.0079774

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*(4), 437–447. doi:10.1016/j.jml.2009.01.004

Roediger, H. L., III, & Butler, A. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20–27. doi:10.1016/j.tics.2010.09.003

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-Enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. doi:10.1111/j.1467-9280.2006.01693.x

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463. doi:10.1037/a0037559

Storm, B. C., Bjork, R. A., & Storm, J. C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. *Memory & Cognition, 38*(2), 244–253. doi:10.3758/MC.38.2.244

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. doi:10.1111/j.1467-9280.1992.tb00036.x

