Educ Psychol Rev (2012) 24:419–435
DOI 10.1007/s10648-012-9203-1

REVIEW ARTICLE

When Is Practice Testing Most Effective for Improving the Durability and Efficiency of Student Learning?

Katherine A. Rawson · John Dunlosky

Published online: 31 July 2012

Abstract Although summative testing is often maligned within educational communities, practice testing is one of the most well-established strategies for improving student learning. Despite the wealth of empirical evidence that testing can enhance learning, teachers and students underutilize practice testing as a learning strategy. Accordingly, a high-level goal of this paper is to advocate for increased use of practice testing as a means for improving student learning. To this end, we discuss prior research establishing the generality of test-enhanced learning as well as prior research that points to conditions under which practice testing is particularly effective. We then summarize some recent research that explores schedules of practice testing that will not only produce durable learning, but will do so most efficiently. To briefly foreshadow, a particularly effective schedule involves practicing retrieval until target information is correctly recalled once during initial learning and then relearned to one correct recall in three to four subsequent sessions. Finally, we argue that exploring both criteria—durability *and* efficiency—can be valuable for evaluating the utility of learning techniques, and we offer some basic prescriptive conclusions for students and educators as well as recommendations for future research.

Keywords Test-enhanced learning · Retrieval practice · Testing effects · Relearning · Long-term retention · Efficiency

The use of testing in formal education has increasingly attracted negative attention and debate within educational communities, particularly concerning high-stakes summative assessments. As a result, "test" is understandably a negative word to many teachers and students. In light of this view of testing as an enemy of education, our goal is to describe

This work was supported by a Collaborative Award from the James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind and Behavior, and by the Institute of Education Sciences, U.S. Department of Education, through grant #R305A080316 to Kent State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

K. A. Rawson (✉) · J. Dunlosky
Department of Psychology, Kent State University, P.O. Box 5190, Kent, OH 44242-0001, USA
e-mail: krawson1@kent.edu

conditions under which testing can be the friend of education for students attempting to achieve many learning goals.

Although testing is most commonly used as a means to evaluate student learning, what many students and educators may not realize is that testing can also be used to improve student learning, particularly when students are provided with feedback (as will be discussed further below). Indeed, practice testing is one of the most potent strategies for enhancing learning. Over the last 100 years, several hundred experiments have shown positive effects of practice tests on learning and memory, with recent studies even demonstrating the benefits in authentic classroom contexts (e.g., McDaniel *et al*. 2011; Roediger *et al*. 2011). Given the wealth of empirical evidence for test-enhanced learning, it is perhaps surprising that practice testing as a learning strategy receives minimal discussion in educational psychology textbooks. Furthermore, outcomes of recent survey studies suggest that students underuse self-testing during self-regulated learning (e.g., Karpicke *et al*. 2009; Kornell and Bjork 2007).

Accordingly, a major goal of this paper is to advocate for the use of practice testing as a means for improving student learning. To this end, we first briefly discuss prior research establishing conditions under which practice testing is particularly effective. We then summarize recent research from our lab that explores schedules of practice testing that will not only produce durable learning, but will do so most efficiently. Finally, we argue that exploring both criteria—durability *and* efficiency—can be valuable for evaluating the utility of learning techniques, and we offer some basic prescriptive conclusions for students and educators as well as recommendations for future research.

When Is Practice Testing Most Effective?

The influence of practice testing on learning is one of the most well-established effects in all of cognitive psychology (for recent reviews, see Dunlosky *et al*. 2012; Rawson and Dunlosky 2011; Roediger and Butler 2011). Several different methods have been used to reveal test-enhanced learning. In some research, learners are presented with to-be-learned material for initial study followed either by a practice test or no practice test. The modal finding is that performance on a subsequent final test is greater following a practice test versus no practice test (e.g., Butler *et al*. 2007; Chan 2009; Fazio *et al*. 2010; Helder and Shaughnessy 2008). Of course, a practice test usually involves re-exposure to at least some of the to-be-learned information (any information explicitly contained in the question prompt and any target information that is correctly retrieved from memory), which raises concerns that any benefit from a practice test actually reflects an effect of restudying the re-exposed information rather than an effect of testing per se. Accordingly, much of the contemporary research has pitted practice testing against restudy (i.e., re-presentation of the to-be-learned information). Even when compared to this restudy-control group, final test performance is typically greater after practice testing than after restudying (e.g., Carpenter 2009; Carpenter and DeLosh 2006; Morris *et al*. 2005). Finally, although practice testing is consistently more effective than restudy, combining practice tests with restudy is typically more effective than either testing or restudy alone (e.g., Agarwal *et al*. 2008; Butler and Roediger 2008; Carpenter *et al*. 2009; Cull 2000; Pyc and Rawson 2010). Providing feedback or restudy is particularly important when students respond incorrectly on a practice test in that it minimizes or eliminates perseveration errors (e.g., Butler and Roediger 2008), although feedback need not be provided immediately to produce these protective effects (e.g., Metcalfe *et al*. 2009).

Substantial research has also established the generality of the effects of practice testing. Although most of the research has involved college students, testing effects have been shown

with preschoolers, elementary school students, middle school students, and older adults (e.g., Balota *et al*. 2006; Bouwmeester and Verkoeijen 2011; Carpenter *et al*. 2009; Fritz *et al*. 2007). Furthermore, although much of the prior work has examined learning of simple verbal materials (such as lists of words or word pairs), testing has also been shown to enhance learning of more complex verbal materials (such as lengthier text passages; e.g., Agarwal *et al*. 2008; Chan 2009, 2010) and learning of illustrations (such as diagrams and maps; e.g., Carpenter and Pashler 2007; Glover 1989; Rohrer *et al*. 2010). Test-enhanced learning has been shown on many kinds of criterion measures, including those that primarily measure memory for the practiced items and those that primarily measure comprehension or application of the practiced items (e.g., Butler 2010; Johnson and Mayer 2009; McDaniel *et al*. 2009). Furthermore, above and beyond any benefit of practice testing for practiced items, several recent studies have shown benefits for related information that was not tested during practice (e.g., Chan 2009, 2010; Cranney *et al*. 2009). Across the various outcome measures that have been examined, testing effects have been shown to last over moderate delays (e.g., a week) and even up to several weeks or months (e.g., Butler and Roediger 2007; Carpenter *et al.* 2009; McDaniel *et al*. 2007). Indeed, the advantage of practice testing over control conditions often are larger on delayed tests than on immediate tests, with evidence suggesting that testing can retard forgetting (e.g., Johnson and Mayer 2009; Roediger & Karpicke 2006).

In sum, this vast body of research leaves no doubt that practice testing enhances learning. Accordingly, an increasing number of studies have focused on identifying factors that can increase the effectiveness of testing. First, as mentioned above, combining practice tests with restudy is typically more effective than either testing or restudy alone. This advantage of testing plus restudy versus testing alone obtains when information is restudied immediately after the practice test or with some delay between the practice test and restudy.

Second, several studies have directly compared the effectiveness of different kinds of practice tests (e.g., Butler and Roediger 2007; Carpenter and DeLosh 2006; Glover 1989; McDaniel *et al.* 2007). Results suggest that practice tests are more effective when they require retrieval of information from long-term memory (e.g., such as a short answer question that presents learners with a key term and asks them to recall the definition from memory) rather than recognition-based tests (e.g., a multiple-choice question that presents learners with a key term and asks them to select the correct definition from among a list of alternatives). Furthermore, the advantage of retrieval-based practice tests over recognition-based tests has been shown even when the final test is recognition-based.

Third, several studies have shown that increasing the number of practice tests leads to greater levels of performance on subsequent criterion measures (e.g., Cull *et al*. 1996; Glover 1989; Pavlik and Anderson 2005; Vaughn and Rawson 2011). However, the benefit of multiple practice tests depends considerably on the timing of those tests. Practice tests that are spread out with longer intervals between each next test are much more effective than practice tests that are completed in close succession. The advantage of longer versus shorter intervals between practice tests holds for the timing of repeated tests within a session (e.g., Cull *et al.* 1996; Karpicke and Roediger 2007; Pashler *et al*. 2003; Pyc and Rawson 2009) as well as for the time intervals between learning sessions (e.g., Bahrick 1979; Cepeda *et al*. 2008).

In sum, the prevailing evidence supports the prescriptive conclusion that a particularly effective way to enhance student learning is to engage in repeated, retrieval-based practice tests that are followed by restudy and that are distributed across time (hereafter referred to as *distributed test–restudy* for sake of brevity).

Optimizing Schedules of Retrieval Practice: How Much Is Enough?

Distributed test–restudy practice is highly effective for enhancing the durability of learning. To be clear, by "durable" we are referring not to a kind of learning (e.g., rote versus meaningful) but rather to the duration over which learning is maintained. As noted above, prior research has shown that the benefits of practice testing can persist across days, weeks, and even months. However, although durable learning is certainly an important goal, educators and students are constrained by the amount of time and effort that may be spent on learning a given set of material. Thus, researchers interested in facilitating the real-world application of test-enhanced learning (or any other study technique, for that matter) in authentic educational contexts must consider not only the conditions under which practice testing is most effective but also the schedules of practice testing that most efficiently achieve high levels of durable learning. The practical question can be stated simply: Distributed test–restudy is good, but how much is enough? To this end, we next summarize some of our recent work that further reveals the potency of practice testing and highlights how it can be most effectively used to achieve durable and efficient learning.

Before we discuss the outcomes of particular experiments, we briefly overview some key aspects of methodology that are common across the studies that we summarize below. First, in all of these studies, the to-be-learned information included definitions of key concepts from course-relevant material. Key concepts are the central component of most textbook chapters in many content domains (e.g., consider the list of key terms found at the end of most textbook chapters), with much of the additional information within a chapter intended to elaborate and illustrate those concepts. Likewise, many classroom activities are focused on helping students learn about key concepts. Key concepts are an important part of the foundational knowledge a student must acquire to master a topic, and they provide the conceptual building blocks for more advanced coursework. Thus, helping students learn these concepts has the potential for improving student achievement.

Second, all of the schedules of practice testing that we describe below involved the conditions that have already been shown to be particularly advantageous (as summarized above)—namely, all schedules included repeated practice tests, all practice tests were retrieval-based, and all tests were followed by restudy opportunities. Whereas these conditions are held constant, the key manipulations concerned the amount and timing of practice. Third, regarding the amount of practice, prior research provides some relevant evidence suggesting that the benefit of retrieval practice is particularly pronounced when students answer correctly versus incorrectly (e.g., Pashler *et al.* 2003; Kang *et al.* 2011).[1] Additionally, a recent survey found that when students engage in self-testing via the use of flashcards during self-regulated learning, 65 % of the students reported practicing until information was correctly recalled at least once Wissman *et al.* (2012).

Accordingly, all studies described below involved *learning to criterion*, such that concepts were presented for as many practice tests as needed until a student could correctly recall the target definition. The key manipulations concerned (a) the number of times a definition was correctly recalled before practice with that concept was terminated and (b) the timing of those correct retrievals. Thus, our answer to the critical question about how much

[1] Concerning the consequences of answering incorrectly on a practice test, most evidence suggests that as long as students receive feedback that includes the correct answer, failed tests have minimal negative effects (e.g., Butler and Roediger 2008; Kang *et al.* 2011). Research has even shown that retrieval failures followed by feedback can promote memory under some conditions (Kornell *et al.*. 2009; Richland *et al.* 2009; Vaughn and Rawson 2012). However, the benefits of practice testing are strongest when retrieval is successful.

practice is enough is cast in terms of the number of correct retrievals, not the number of trials or tests, and the starting point for our answer about the optimal amount of distributed test–restudy practice is "at least until you get it right." The trickier part of the answer will concern how many times, and when.

How schedules of distributed test–restudy can influence the durability and efficiency of learning is well illustrated by a recent large-scale study involving 335 undergraduate students enrolled in an Introductory Psychology course (Rawson and Dunlosky 2011, Experiment 3). Materials included modified excerpts from Introductory Psychology textbooks on topics commonly taught in this class (including visual perception, the nervous system, memory, and social judgments). Each textbook excerpt included several key concepts and their definitions; examples of these concepts are provided at the top of Table 1. Within the context of the experiment, each participant was assigned to learn two of the four concept sets. The other two concept sets were not presented to that participant during any of the learning sessions of the experiment, but the concepts were included on the final tests. The purpose of this manipulation was to evaluate how well students learned the target concepts via exposure to these concepts in their class, outside of the context of the experiment (we refer to this condition as *baseline control*). Assignment of sets to practice versus baseline control conditions was counterbalanced across participants.

During the initial learning session, students began by studying the textbook excerpts. The target concepts and their definitions were then presented via computer one at a time in isolation for additional study. After this initial study phase, the practice phase began. During the practice phase, the computer presented each concept for a practice cued recall test in which the concept was used as a cue (e.g., "What is the just-world hypothesis?"), and students were prompted to type in as much of the definition as they could remember. After completing their response, the computer presented the correct answer for students to compare to their own answer and to restudy. On trials in which the student's response was not correct, the concept was placed at the end of the list for another practice test–restudy trial later (analogous to putting a flashcard at the back of the stack).

Table 1 Examples of key concepts used in our research

Concepts from College-Level Introductory Psychology
Topic: Social Judgments
The *just-world hypothesis* refers to the strong desire or need people have to believe that the world is an orderly, predictable, and just place, where people get what they deserve.
Topic: Memory
Episodic memory is memory for personally experienced events along with information about the time and context in which they occurred.
Topic: Visual Perception
Shape constancy is the perception that an object remains the same shape even though the retinal image changes when its orientation to us changes.
Concepts from Middle School Courses in Math and Science
Topic: Statistics
The *mode* is the number or object that appears most frequently in a set of numbers or objects.
Topic: Genetics
An *allele* is one member of a pair of genes that is located at the same position on a specific chromosome.
Topic: Geography
Erosion is the process in which the earth is worn away and taken from one place to another one.

Each concept continued to be presented for practice test–restudy trials until a student correctly recalled the definition at least once.[2] For half of the students, a concept was dropped from further practice in the initial learning session after the definition was correctly recalled once. For the other half of the students, each concept continued to be presented for additional practice test–restudy trials until the definition was correctly recalled three times. We refer to this manipulation as *initial learning criterion* (one versus three correct recalls).

After the initial learning session was completed, students returned to the lab 2 days later to complete a *relearning* session. The relearning session was similar to the initial learning session, except concepts were not presented for initial study. Rather, the session immediately began with a practice test followed by restudy for each concept. Concepts that were not correctly recalled on this first practice test were presented again later for another practice test–restudy trial and so on, until each definition was correctly recalled once. A concept was removed from further practice in the relearning session after one correct recall. Thus, the criterion within a relearning session was always the same for all concepts and students. However, we did manipulate the number of relearning sessions each student completed. One group of students only completed one relearning session. Another group of students returned to the lab to complete a second relearning session, and other groups of students completed a third, fourth, or fifth relearning session. We refer to this manipulation as *relearning level* (one, two, three, four, or five relearning sessions).

Because we are interested in exploring how to promote durable learning, students completed a final test 1 month after their last relearning session and another final test 4 months after their last relearning session. As mentioned above, each of these tests included the concepts that students had practiced during the experiment as well as the baseline control concepts that they had not practiced experimentally. For all concepts, the test involved cued recall of the definitions, as during practice.

Initial learning: how much is enough? Given that the design of this experiment is somewhat complex, we will simplify our summarization of the results by first describing the key outcomes associated with the effects of initial learning criterion and then describing the key outcomes associated with the effects of relearning level. Concerning the effects of initial learning criterion, key outcomes are presented in Fig. 1. To revisit, students practiced concepts until definitions were correctly recalled either one or three times during initial learning. All students returned 2 days later to complete a relearning session that began with a practice test trial for each concept. Note that this first practice test also functionally serves as an interim test, indicating retention of initial learning across a 2-day delay (see the bars labeled "interim test after initial learning" in Fig. 1). Performance on this interim test was significantly greater for students who had practiced to three correct recalls versus one correct recall during initial learning (73 versus 58 %). At this point, the additional time during initial learning to achieve a higher criterion level appears to have been time well spent.

However, the effect of initial learning criterion was markedly attenuated by subsequent relearning. Consider the bars labeled "interim test after some relearning" in Fig. 1. These

[2] A definition did not need to be recalled verbatim to be counted as correct. Students could provide an answer using their own words, as long as the response captured the correct meaning of the definition. The decision about whether a response was correct was not based on computer scoring, because a reliable automated system for scoring the meaning of sentence-length responses is not readily available. Rather, the correctness of a response was based on self-scoring judgments that students were prompted to make by comparing their own response to the correct definition. With appropriate support, students can make these judgments in a highly accurate manner. For details about this judgment procedure, see Rawson and Dunlosky (2011). To foreshadow, we will briefly revisit this issue in the General Discussion: Conclusions and Recommendations.

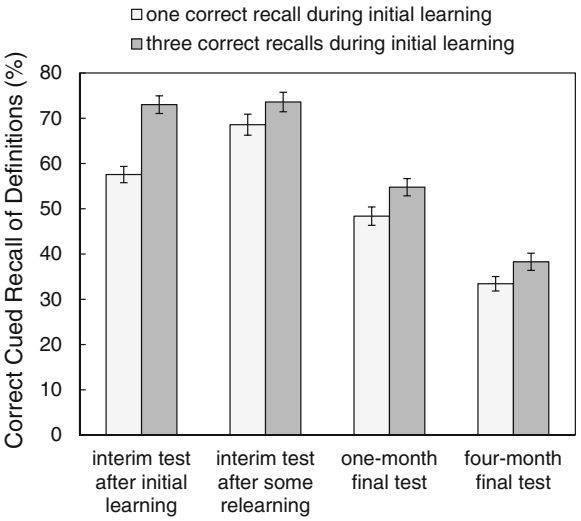

Fig. 1 Mean correct cued recall of definitions as a function of initial learning criterion (one versus three correct recalls during the initial learning session). Cued recall performance is plotted for four time points: interim test after initial learning (i.e., performance on the first practice test given at the beginning of the first relearning session), interim test after some relearning (i.e., performance on the first practice test given at the beginning of the last relearning session; see text for details), and final tests administered 1 and 4 months after the last relearning session. *Error bars* report standard error of the mean

bars report performance on the first practice test trial administered in the last relearning session for students in the groups who were assigned to complete two to five relearning sessions. Given that this test was completed at the beginning of the last relearning session, students had obviously not yet completed all scheduled relearning at this point. However, students had completed at least one prior relearning session (e.g., students who were assigned to the group completing four relearning sessions had already completed three prior relearning sessions at this point). What is clear from inspection of performance at this intermediate time point is that in contrast to the sizeable benefit of a higher criterion level during initial learning that was evident prior to relearning, a higher initial criterion no longer produced much of an advantage after some relearning had taken place. Note that the attenuation of the effect was largely due to the lower criterion group catching up to the higher criterion group with relearning. Likewise, the two sets of bars on the right side of Fig. 1 reveal similar outcomes on the final tests that took place 1 and 4 months after all relearning had been completed.

A key conclusion here is that as long as concepts are practiced until their definition is correctly retrieved at least once during initial learning, the amount of additional practice during initial learning does not seem to matter much if students are subsequently going to engage in relearning. Furthermore, although a small effect of initial learning criterion persisted across relearning in this experiment, the difference was not statistically significant when controlling for minor differences in baseline performance. Furthermore, the effect was not significant in the other two experiments reported in Rawson and Dunlosky (2011) or in our follow-up research (Rawson and Dunlosky 2012; Rawson and Dunlosky, unpublished).

What about the overall efficiency associated with higher initial criterion levels? On one hand, just as relearning reduces the benefit of a higher initial criterion, so too does relearning reduce the overall cost of achieving a higher initial criterion. Not surprisingly, to achieve a

criterion of three correct recalls versus one correct recall during initial learning, students required about two additional practice trials per concept during the initial learning session, which translated into about an extra 2 min per concept (see Fig. 2). However, as is apparent from inspection of Fig. 2, some of this additional cost was recouped with faster relearning in subsequent sessions. For example, students who practiced concepts to three correct recalls during initial learning only needed about 1.6 min per concept to reach criterion in the first relearning session, whereas students who stopped after one correct recall during initial learning needed 2.1 min per concept to reach criterion in the first relearning session. On the other hand, the cost associated with a higher initial criterion was not completely recouped by faster subsequent relearning. Across all sessions, students ended up spending close to one additional minute per concept overall when concepts were practiced to three versus one correct recalls during initial learning (9.0 versus 8.1 min per concept, respectively).

Relearning: bang for the buck In contrast to the limited benefit of a higher initial learning criterion, the benefit of increasing the number of relearning sessions is substantial. To illustrate, consider performance on the final tests administered 1 and 4 months after the last relearning session (see Fig. 3). Completing three versus one relearning session produced a 60 % relative increase in performance on the 1-month test (by comparison, completing three versus one correct recall during initial learning only produced a relative increase of 13 %). Completing five versus one relearning sessions produced an 86 % relative increase on the 1-month test and a 64 % increase on the 4-month test. Furthermore, the absolute level of performance achieved on these tests is impressive, given the difficulty of the test format (requiring recall of definitions) and the lengthy retention interval. As an additional point of comparison, consider performance for the baseline control concepts (the two sets of concepts for each student that were not presented during any learning session but were included on the final tests). Despite the fact that students reported having encountered 54 % of these concepts in their class, performance across the two final tests was only 11 % for these non-practiced control concepts. Thus, relearning produced marked improvements over students' "business as usual" approach to learning.

The benefit of increased relearning is substantial, but what are the costs associated with relearning? Unavoidably, each additional relearning session requires additional time and

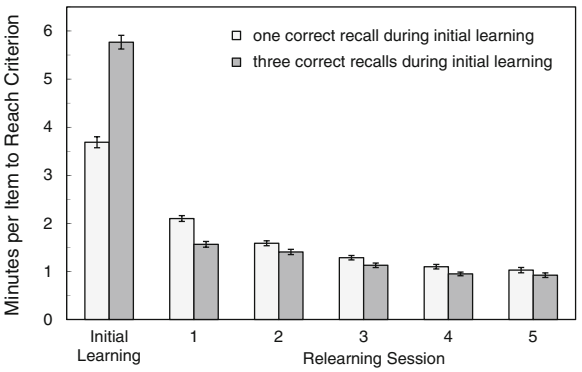

Fig. 2 Mean number of minutes needed per concept for students to reach the assigned criterion level in a given session (either one or three correct recalls during *Initial Learning*, and one correct recall in all relearning sessions). *Error bars* report standard error of the mean

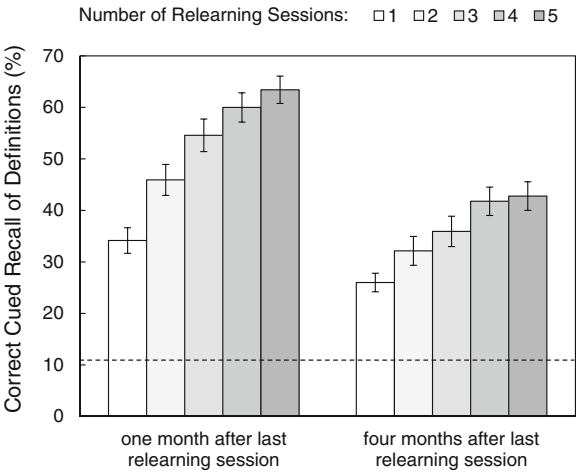

Fig. 3 Mean correct cued recall of definitions as a function of the number of relearning sessions completed prior to the final test that took place either 1 or 4 months later. The *dotted line* represents performance for course-relevant concepts assigned to the baseline control condition (i.e., that were not practiced in the context of the experiment). *Error bars* report standard error of the mean

effort. The good news is that the amount of additional time needed declines across relearning sessions. For example, students needed 1.8 min per concept to achieve one correct recall during the first relearning session, but they needed less than 1 min per concept to achieve one correct recall in the fifth relearning session. Figure 4 reports the total amount of time spent per concept across all learning sessions, which includes initial learning and all of the relearning sessions completed by each group. When considered in light of the levels of long-term learning that are achieved (as shown in Fig. 3), the relatively modest amount of additional time spent on a third or fourth relearning session appears to have been a particularly good investment.

Furthermore, some of the additional cost from increasing the number of relearning sessions is recouped by faster reacquisition of forgotten information *after* the final test.

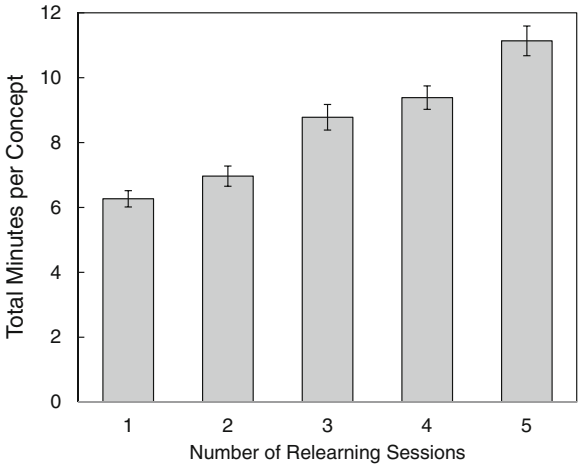

Fig. 4 Mean number of minutes spent per concept in total across all learning sessions prior to the 1-month test, including during initial learning and all relearning sessions completed by each group. *Error bars* report standard error of the mean

Although we did not mention this aspect of the method in our summary above, after each of the final cued recall tests, the concepts were again presented for test–restudy trials until they were relearned to a criterion of one correct recall. As Nelson (1985) noted, "Educators generally realize that most of what they teach their students does not remain recallable, or even recognizable, for extended periods of time. Rather the hope is that their teaching will have been beneficial because the student could relearn the information relatively quickly" (p. 478). Indeed, increasing the number of relearning sessions facilitated rapid reacquisition. After the 1-month test, students who only had completed one prior relearning session needed 1.7 min per concept to achieve one correct recall, whereas students who had completed four or five prior relearning sessions only needed around 1.1 min to achieve one correct recall. Thus, relearning not only enhances retention, it facilitates reacquisition of forgotten information.

As one last approach to putting the cost/benefit analysis of relearning into perspective, we can compare outcomes for students who practiced concepts until correctly recalled three times during initial learning and then completed one relearning session to outcomes for students who practiced concepts until correctly recalled once during initial learning and then completed three relearning sessions (for brevity, we refer to these as the 3–1 and 1–3 groups, respectively). Note that students in both groups experienced the same level of retrieval success during practice (four correct recalls across learning sessions). Adding up the total amount of time spent across learning sessions, the 1–3 schedule only required around two extra minutes per concept compared to the 3–1 schedule, but the 1–3 schedule yielded significantly higher performance on the 1-month test (53 versus 40 %) and faster reacquisition (1.3 versus 1.7 min per concept). Thus, relearning provides relatively good bang for the buck.

Generality of the effects of relearning In addition to the results from the large-scale study summarized above, outcomes from several additional experiments provide further evidence for the substantial benefits of relearning. We first briefly summarize outcomes from a recent experiment involving middle school students. We then end this section with a description of an on-going project that evaluates the degree to which retrieval practice spaced across sessions can improve actual course grades for students enrolled in an Introductory Psychology course.

Potency of retrieval practice for middle school students Given the volume of research on the benefits of retrieval practice for enhancing learning, surprisingly few studies have involved learners younger than college-age students, and most of these have involved learning simple verbal materials such as word lists. Thus, establishing the potency of distributed retrieval practice for younger students' learning of more complex material is important. In a recent experiment (Rawson, Wissman and Dunlosky, unpublished) we worked with 8th graders who were studying key concepts from a unit on statistics (for an example concept from the unit used in this study as well as examples from other units that we have used in research with middle school students, see Table 1). For each student, half of the concepts were practiced and the other half were assigned to the baseline control condition in which concepts were not practiced in the experiment but were included on the final test. All students completed an initial learning session in which concepts were presented for practice test–restudy trials until each one was correctly recalled once. As in the research with college students described above, on each trial the concept was presented as a cue, students typed in as much of the definition as they could remember, and students were then shown the correct answer to compare to their own answer and to restudy. After initial learning, students

subsequently completed zero, one, two, or three relearning sessions, with each session separated by a few days. As during initial learning, each relearning session involved practice test–restudy trials until each definition was correctly recalled once.

One month after the final session (either after the initial learning session for the zero-relearning group or after the last relearning session in each of the other three groups), students completed a final cued recall test over the practiced concepts as well as the baseline control concepts. Overall, final test performance was considerably greater for practiced concepts than for baseline control concepts. Most important, the size of the advantage for practiced items over baseline control increased monotonically with the number of relearning sessions (e.g., the Cohen's *d* effect sizes associated with zero versus three relearning sessions were 0.6 versus 1.7, respectively). Of course, each relearning session necessarily required additional time. But once again, the amount of additional time needed declined across relearning sessions. For example, students spent about 3 min per concept to achieve one correct recall during the initial learning session, whereas they only needed around 1 min per concept to achieve one correct recall in a third relearning session. Thus, for middle school students as well as college students, the costs associated with relearning are relatively modest when considered in light of the marked benefits of distributed retrieval practice for improving durable learning.

Potency of retrieval practice within an Introductory Psychology course To conclude this section, we point to initial results from a series of experiments currently in progress. In all of our studies summarized above, the outcome measures have included cued recall tests administered weeks or months after practice, given the educational goal of equipping students with durable foundational knowledge that in turn can foster learning in advanced coursework and real-world application. However, students understandably focus heavily on achieving the relatively short-term goal of doing well on course exams. To what extent can the practice schedules described above also achieve the short-term goal of improving outcomes on course exams?

Our ongoing line of research is designed to address this issue. In brief, an instructor who teaches two large sections of Introductory Psychology provided us with sets of key concepts that she would be covering in her classes, and students enrolled in her classes were invited to participate in the study. Students completed practice sessions in the lab that involved learning each concept to criterion during initial learning and then relearning the concepts in three subsequent sessions. In the context of the experiment, students practiced half of the course concepts, whereas the other half were not presented for practice (assignment of concept sets to practice versus baseline control was counterbalanced across participants). The practice sessions were scheduled to align with the instructor's lectures, such that the concepts introduced for practice each week were the same concepts the instructor was presenting in class that week. Students granted permission for the instructor to release their exam scores, so we were able to examine performance on the sections of the exams relevant to the concept sets that were included in the experiment. Early results are promising—on average, students correctly answered 83 % of the questions tapping concepts they practiced in the experiment, versus only 70 % for questions tapping the set of concepts that they did not practice in the experiment (i.e., reflecting how well they learned concepts on their own). This initial outcome is particularly impressive given the differences between the kind of practice test used in the experiment (cued recall of definitions) and the questions used on the course exams (multiple-choice tests that often tapped comprehension or application of the concepts). We have recently

completed data collection for a follow-up experiment intended to replicate and extend these outcomes, but early results suggest that distributed test–restudy involving relearning will be effective for satisfying both the short-term goal of promoting classroom performance as well as the long-term goal of promoting durable knowledge.

General Discussion: Conclusions and Recommendations

We began our review by pointing to the incontrovertible fact that practice testing is a potent strategy for improving student learning, particularly when it involves retrieval-based tests with restudy opportunities that are distributed across time. The wealth of research that firmly establishes these conclusions provides a solid foundation from which to explore how to optimize schedules of distributed test–restudy practice. Accordingly, much of our recent work has been directed at the question: How much is enough?

Most generally, our answer to this question is framed in terms of bang for the buck. Our emphasis on both the benefits and the costs associated with practice schedules is motivated by the competing demands that are present in most educational contexts: Students are expected to learn substantial amounts of knowledge and numerous skills within and across classes, but teachers have a limited amount of classroom time to spend on instruction, and students have a limited amount of time and effort to spend on studying. So, although we know that practice testing works, we argue that the practical goal for future research should be twofold: to identify schedules of testing that lead to durable learning and to identify when time spent on additional testing is time well spent.

To that end, we offer some tentative prescriptive conclusions based on outcomes from the research we have conducted so far (some of which has been summarized above). We offer prescriptions for initial learning and for relearning in turn, and for each we will make both qualitative and quantitative recommendations. Concerning initial learning, our strongest conclusion is that students should at least practice until target information is correctly recalled once. Concerning more precise quantitative conclusions about the optimal amount of practice beyond the first correct recall, unfortunately, no clear conclusion emerges at this point. At most, we would recommend that teachers and students consider the following factors when making decisions about how to allocate study time. First, the advantages of practicing to a higher initial criterion (e.g., three correct recalls) are that (1) if a student fails to engage in enough (or any) subsequent relearning, a higher initial criterion provides some protection against poor retention, and (2) if the student does engage in subsequent relearning, some of the initial cost in time to achieve a higher criterion will be recouped by faster relearning. However, the disadvantages of practicing to a higher initial criterion are that (1) the amount of additional time spent during initial learning is not completely recouped by relearning, and (2) a persistent performance advantage is not guaranteed and will likely be small at best. Additionally, although we have not empirically evaluated this possibility, our speculation is that other potential costs of continuing practice to achieve a higher criterion during initial learning might include boredom, fatigue, or frustration, which would be counterproductive if students were then less motivated to use practice testing as a learning strategy subsequently.

Concerning relearning, the available evidence supports stronger conclusions. Put simply, relearning pays big dividends for the investment of time. Engaging in multiple relearning sessions yields sizeable benefits to both short-term and long-term retention of learned information. Additionally, completing more relearning sessions produces diminishing incremental costs (i.e., in time spent studying), because the speed of relearning accelerates across

sessions. Indeed, in our studies, students are frequently able to correctly recall most of the concept definitions on their first attempt in later relearning sessions. Moreover, our speculation is that relearning sessions are unlikely to produce undue boredom, fatigue, or frustration, given that the majority of students in our studies exhibit rapid relearning (some relearning sessions take 5 min or less). Students may even be more motivated to use practice testing as a learning strategy due to the experience of relatively quick and easy success. With that said, certainly some point will be reached at which additional relearning sessions no longer produce meaningful improvements in durable learning. The results shown in Fig. 3 suggest as much, given the minimal differences in long-term retention after four versus five relearning sessions. However, other recent data from our lab suggests that the potency of additional relearning sessions can be restored by increasing the interval between sessions. Thus, a promising direction for future research will be to explore how to optimize the timing of relearning sessions to further increase the benefits while minimizing increased costs in time used for learning. Nonetheless, the clear take-home message is that multiple relearning sessions are powerful with respect to enhancing the durability of learning and at the same time relatively cheap with respect to the efficiency of learning.

We are reluctant to declare a universal 'magic number' of relearning sessions at this point, which at a minimum would require much more research to establish the generality of the patterns we have observed in these and other studies. However, our tentative recommendation is that students engage in at least three relearning sessions, because the benefits consistently appear to warrant the costs. With that said, note that our research has involved undergraduate students and middle school students and primarily has focused on supporting learning of key concepts from course materials. The number of relearning sessions at which asymptotic levels of performance are observed may differ for younger or older learners, or for other kinds of to-be-learned material. The quantitative effects of number of relearning sessions may also differ depending on the retention interval or the kind of criterion test used to measure performance. These possibilities suggest other important directions for future research.

Finally, we address one additional issue of importance for effectively implementing distributed test–restudy practice for key concepts of the sort involved in the research described here. Up to this point, we have been largely silent about how one determines when the definition of a concept has been correctly recalled. For word-length responses, the decision is trivially easy for both students and computers. For example, suppose a student is attempting to learn foreign language vocabulary and studies the translation *maison-house*. If the student is prompted with *maison—???* on a subsequent practice test and incorrectly responds with the word *spoon*, as long as the correct answer is made available for comparison, it would be easy for the student (or the computer, if automated scoring is involved) to recognize that the answer is not correct and that the item will require additional practice.

In contrast, for materials that require more complex responses such as the ones used in the research described here, evaluating the accuracy of a response is much less straightforward. Whereas substantial research has shown that computer algorithms can reliably score the semantic content of lengthier responses (i.e., one or more paragraphs; Landauer *et al*. 2007), the extent to which computer algorithms can reliably score the meaning of a sentence-length response (i.e., the length of definitions for most key concepts) is much less well established (for progress on this front, see Kintsch and Mangalath 2011; McCarthy *et al*. 2009). Given that it is not feasible to have teachers grade students' responses during learning sessions, the only remaining option is to have students evaluate the quality of their own responses. The key here is

that to effectively implement schedules of distributed test–restudy that involve learning to criterion, students must be able to accurately judge when their response is correct. If students are overconfident and indicate that they are correctly recalling a definition when in fact it is incorrect, the benefit of using retrieval practice to learn to criterion will be attenuated, because students will not actually be achieving the sought-after criterion (i.e., concepts will be dropped from practice before they have been correctly recalled; see Dunlosky and Rawson 2012). Thus, it is essential that students not be overconfident when evaluating their recall responses.

Without any form of feedback, students are often substantially overconfident when they evaluate their responses. Our solution to this problem involved developing and evaluating the efficacy of different forms of feedback to help students accurately evaluate their recall responses. Fortunately, feedback in the form of idea units (i.e., breaking the correct definition into the smaller constituent ideas that are required to receive full credit) that students can check against their own responses largely diminishes overconfidence and supports highly accurate evaluations (for reviews, see Dunlosky and Lipko 2007; Rawson and Dunlosky 2012; Rawson and Dunlosky, unpublished). In the experiments described above, students made these idea-unit judgments immediately after each retrieval attempt, and these judgments were used by the computer to decide when a given concept had met the appropriate learning criterion. With minimal training, students can generate idea units and use them to accurately evaluate the quality of their recall on their own (Dunlosky *et al*. 2011), so the use of idea-unit judgments to promote accurate self-evaluation during retrieval practice for key concepts can be used broadly without any technological support.

The inherent problem of identifying a reliable means to score the accuracy of complex responses (e.g., concept definitions) in real time is perhaps one reason why minimal prior research has examined the effects of criterion learning and relearning. Indeed, one important way in which our work extends beyond virtually all prior research on testing effects is that it involves manipulations of initial learning criterion and of relearning level. Our initial investigations of these variables provide a foundation for further research to explore how to optimize schedules of distributed test–restudy to maximize not only the durability of learning but also the efficiency of learning (an equally important but usually overlooked goal).

Nevertheless, the implementation of practice testing in the classroom and in students' self-regulated learning need not await further research for students to realize benefits from this highly potent strategy. Even small amounts of practice testing that do not necessarily involve criterion learning can yield meaningful gains in learning for a relatively small expenditure in time. For example, Daniel and Broida (2004) reported appreciable improvements in performance on course exams for classes involving one quiz per week versus no weekly quiz. McDaniel *et al.* (2011) found that unit exam performance was 17 % greater following even just one low-stakes review quiz prior to the exam. Roediger *et al.* (2011) found similar effects with younger learners. Across experiments, sixth-grade students consistently showed improvements in performance on chapter exams for facts that had been quizzed previously versus not quizzed, and the advantage persisted even on surprise exams administered at the end of the semester 1–3 months later. Finally, students who self-report spontaneously using practice testing outperform students who do not (Gurung 2005; Hartwig and Dunlosky 2012). Thus, we conclude with our strongest and highest-level prescriptive conclusion: Teachers and students should more regularly put practice testing to work in the classroom and in self-regulated learning.

References

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

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

Balota, D. A., Duchek, J. M., Sergent-Marshall, S. D., & Roediger, H. L., III. (2006). Does expanding retrieval produce benefits over equal-interval spacing? Explorations of spacing effects in healthy aging and early stage Alzheimer's disease. *Psychology and Aging, 21*, 19–31.

Bouwmeester, S., & Verkoeijen, P. P. J. L. (2011). Why do some children benefit more from testing than others? Gist trace processing to explain the testing effect. *Journal of Memory and Lanugage, 65*, 32–41.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1118–1133.

Butler, A. C., & Roediger, H. L. I. I. I. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*, 514–527.

Butler, A. C., & Roediger, H. L. I. I. I. (2008). Feedback enhances the positive effects and reduces the negative effects multiple-choice testing. *Memory & Cognition, 36*, 604–616.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. I. I. I. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology. Applied, 13*, 273–281.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1563–1569.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*, 268–276.

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review, 14*, 474–478.

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology, 23*, 760–771.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19*, 1095–1102.

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61*, 153–170.

Chan, J. C. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory, 18*, 49–57.

Cranney, J., Ahn, M., McKinnon, R., Morris, S., & Watts, K. (2009). The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *European Journal of Cognitive Psychology, 21*, 919–940.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14*, 215–235.

Cull, W. L., Shaughnessy, J. J., & Zechmeister, E. B. (1996). Expanding understanding of the expanding-pattern-of-retrieval mnemonic: Toward confidence in applicability. *Journal of Experimental Psychology. Applied, 2*, 365–378.

Daniel, D. B., & Broida, J. (2004). Using web-based quizzing to improve exam performance: Lessons learned. *Teaching of Psychology, 31*, 207–208.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16*, 228–232.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22*, 271–280.

Dunlosky, J., Hartwig, M. K., Rawson, K. A., & Lipko, A. R. (2011). Improving college students' evaluation of text learning using idea-unit standards. *Quarterly Journal of Experimental Psychology, 64*, 467–484.

Dunlosky, J., Rawson, K.A., Marsh, E.J., Nathan, M.J., & Willingham, D.T. (2012). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest.* (in press).

Fazio, L. K., Agarwal, P. K., Marsh, E. J., & Roediger, H. L. I. I. I. (2010). Memorial consequences of multiple-choice testing on immediate and delayed tests. *Memory & Cognition, 38*, 407–418.

Fritz, C. O., Morris, P. E., Nolan, D., & Singleton, J. (2007). Expanding retrieval practice: An effective aid to preschool children's learning. *Quarterly Journal of Experimental Psychology, 60*, 991–1004.

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

Gurung, R. A. R. (2005). How do students really study (and does it matter)? *Teaching of Psychology, 32*, 239–241.

Hartwig, M. K., & Dunlosky, J. D. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review, 19*, 126–134.

Helder, E., & Shaughnessy, J. J. (2008). Retrieval opportunities while multitasking improve name recall. *Memory, 16*, 896–909.

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101*, 621–629.

Kang, S. H. K., Pashler, H., Cepeda, N. J., Rohrer, D., Carpenter, S. K., & Mozer, M. C. (2011). Does incorrect guessing impair fact learning? *Journal of Educational Psychology, 103*, 48–59.

Karpicke, J. D., & Roediger, H. L. I. I. I. (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*, 704–719.

Karpicke, J. D., Butler, A. C., & Roediger, H. L. I. I. I. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory, 17*, 471–479.

Kintsch, W., & Mangalath, P. (2011). The construction of meaning. *Topics in Cognitive Science, 3*, 346–370.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*, 219–224.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 989–998.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis*. Mahwah: Erlbaum.

McCarthy, P. M., Guess, R. H., & McNamara, D. S. (2009). The components of paraphrase evaluations. *Behavior Research Methods, 41*, 682–690.

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., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science, 20*, 516–522.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. I. I. I. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103*, 399–414.

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children‘s and adults’ vocabulary learning. *Memory & Cognition, 37*, 1077–1087.

Morris, P. E., Fritz, C. O., Jackson, L., Nichol, E., & Roberts, E. (2005). Strategies for learning proper names: Expanding retrieval practice, meaning, and imagery. *Applied Cognitive Psychology, 19*, 779–798.

Nelson, T. O. (1985). Ebbinghaus’s contribution to the measurement of retention: Savings during relearning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 472–479.

Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1051–1057.

Pavlik, P. I., Jr., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science, 29*, 559–586.

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*, 437–447.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*, 335.

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology. General, 140*, 283–302.

Rawson, K.A., & Dunlosky, J. (2012). Retrieval-monitoring-feedback (RMF) technique for producing efficient and durable student learning. To appear in R. Azevedo & V. Aleven (Eds.), *International Handbook of Metacognition and Learning Technologies* (in press).

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology. Applied, 15*, 243–257.

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

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

Roediger, H. L., III, Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology. Applied, 17*, 382–395.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 233–239.

Vaughn, K. E., & Rawson, K. A. (2011). Diagnosing criterion level effects on memory: What aspects of memory are enhanced by repeated retrieval? *Psychological Science, 22*, 1127–1131.

Vaughn, K.E., & Rawson, K.A. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin & Review*. doi:10.3758/s13423-012-0276-0.

Wissman, K.T., Rawson, K.A., & Pyc, M.A. (2012). How and when do students use flashcards? *Memory*. doi:10.1080/09658211.2012.687052.