ANNUAL REVIEWS

Annual Review of Psychology

Practicing Retrieval Facilitates Learning

Kathleen B. McDermott

Department of Psychological and Brain Sciences, Washington University in St. Louis, St. Louis, Missouri 63130, USA; email: kathleen.mcdermott@wustl.edu

Annu. Rev. Psychol. 2021. 72:609–33

First published as a Review in Advance on October 2, 2020

The *Annual Review of Psychology* is online at psych.annualreviews.org

https://doi.org/10.1146/annurev-psych-010419-051019

Keywords

retrieval practice, testing effect, retrieval-based learning, memory, learning

Abstract

How do we go about learning new information? This article reviews the importance of practicing retrieval of newly experienced information if one wants to be able to retrieve it again in the future. Specifically, practicing retrieval shortly after learning can slow the forgetting process. This benefit can be seen across various material types, and it seems prevalent in all ages and learner abilities and on all types of test. It can also be used to enhance student learning in a classroom setting. I review theoretical understanding of this phenomenon (sometimes referred to as the testing effect or as retrieval-based learning) and consider directions for future research.

Contents

INTRODUCTION 610
CONCEPTUAL FRAMEWORK 610
HISTORICAL PRECEDENTS 611
MODERN INTEREST IN THE POWER OF RETRIEVAL PRACTICE 614
TERMINOLOGY 616
GENERALIZABILITY 617
  Materials 617
  Age Groups 618
  Test Format (and the Importance of Correct-Answer Feedback) 618
  Retention Interval 619
  Individual Differences 620
EXTENSIONS 620
  Enhancing Classroom Learning 620
  Retrieval-Induced Facilitation: Enhancing Memory for Related, Nontested Information 624
THEORIES AND FRAMEWORKS 625
REMAINING PUZZLES: FUTURE DIRECTIONS 628

INTRODUCTION

Consider how one goes about learning to play a musical instrument. One of the first words that comes to mind is "practice." In learning a musical instrument—much like learning to swim or learning to play chess—we take it for granted that we need to practice the new skill until we become facile at its execution.

But how do we think about acquiring new knowledge from a textbook? What kinds of activities facilitate this kind of learning? Perhaps highlighting comes to mind, or repeatedly reading the chapter, or maybe making some notes or an outline. However, if the eventual goal is to be able to retrieve that knowledge from memory, perhaps practicing retrieval of that information would be a better way to learn. Indeed, retrieval practice is one of the most effective ways of solidifying new knowledge, although this fact is underappreciated by most learners (and teachers). The goal of this review is to consider some evidence about the power of retrieval practice to support learning.

CONCEPTUAL FRAMEWORK

In the context of the musical instrument analogy above, the idea that retrieval practice can help establish long-term memories is perhaps unsurprising. But if you ask college students, who are fairly sophisticated learners, how to best learn new information, their answers suggest that they do not appreciate this fact (Blasiman et al. 2017, Karpicke et al. 2009, Kornell & Bjork 2007, McCabe 2011). Instead, retrieval seems to be considered mostly a measurement device (i.e., assessing, but not influencing, learning).

This observation of a limited understanding of the power of retrieval extends to cognitive psychologists, too, in many respects. The popular three-store model, whereby human memory is conceptualized as occurring in three stages—encoding, storage, and retrieval—dates back at least to work by Köhler (1947) and Melton (1963). The three-store framework holds that after we acquire new information, some of it undergoes storage into long-term memory, and then after a while some of this previously encoded information can be retrieved. The stages are considered logically and temporally separate, though. This mindset overlooks the iterative and interactive processes underlying learning (Gardiner 2007). When we encode and then retrieve, the retrieval process results in the memory being solidified such that it remains more accessible over time (Bjork 1975). Much of everyday, real-world human learning is iterative: For example, in a classroom students read before class, attend lectures, take quizzes, review the information, work with the material before a test in various ways, take a unit test, and then take a final exam. The mindset positing that there are discrete encoding and retrieval phases does not describe well this state of affairs, nor does this viewpoint capture the interactive nature of encoding and retrieval, whereby retrieval can be considered a (re)encoding event.

Ironically, much of the literature demonstrating the positive effects of retrieval practice for subsequent retrievability draws upon this discrete-stage experimental methodology. The advantage of this experimental design choice is that it can demonstrate quite cleanly how retrieval is a "memory modifier," to borrow a phrase from Bjork (1975). There is, however, an emerging realization that if we want our research to be better applicable to educational settings, we as researchers need to explicitly address the iterative nature of encoding and retrieval. I expand on this issue below.

HISTORICAL PRECEDENTS

That practicing retrieval can facilitate later retrieval is an observation that predates the experimental study of memory. Sir Francis Bacon, for example, posited that "If you read a piece of text through twenty times, you will not learn it by heart so easily as if you read it ten times while attempting to recite from time to time and consulting the text when your memory fails" [Bacon 2000 (1620), p. 143; see also James 1890, p. 646, for a similar observation, and Abbott 1909 for an early empirical study].

This self-testing approach, whereby one alternates between reading, self-testing, and feedback (in the form of rereading when retrieval attempts break down) has been coined recitation. In an early study demonstrating the power of recitation, Gates (1917) gave a group of children 9 minutes to learn new materials (nonsense syllables or biographies). He manipulated the percent of the time that was spent in recitation (0, 20, 40, 60, or 80% of the 9-minute learning period, with the remaining time being spent reading the material). As shown in **Figure 1**, performance on a

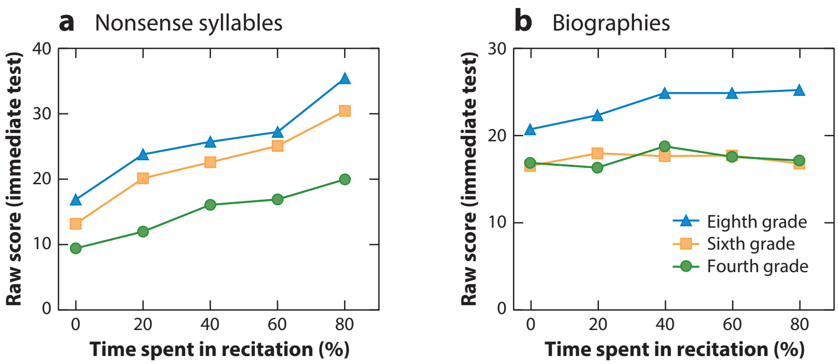

Figure 1

Recitation improves memory. Students who spent a higher percentage of their 9-minute study period in recitation recalled more nonsense syllables (panel *a*). Effects on recall of biographies (panel *b*) were mixed, with improvement being isolated to students in eighth grade. Data from Gates (1917, tables XVII, XXI).

final test (given after the 9-minute period) varied as a function of the amount of recitation, although the effect differed across materials. For nonsense syllables, there was a strong effect, such that the more time spent in iterative retrieval practice with feedback (recitation), the better the recall. For biographies, a small effect emerged for eighth-graders but not for the younger students.

Although recitation provides a compelling example of the power of retrieval practice in facilitating memory, a more surprising phenomenon can be observed when participants are not able to turn back to the original material to restudy it—that is, when participants study the material, take an initial test (or not), and then take a final test. Participants taking an initial test receive no feedback (aside from the subjective sense of how well they are doing), and the comparison (no test) condition receives no further interaction with the material besides the study phase and the final test. Spitzer (1939) provided a large-scale manipulation using this type of design. He recruited 3,605 sixth-grade students to spend 8 minutes reading ∼600-word fact-heavy, age-appropriate articles; the topics were bamboo and peanuts. All students took an initial test (multiple choice) and then took the same test again, which was unexpected. Some students then took an unexpected third test. Spitzer varied the delay at which the initial and subsequent tests occurred. Students were divided into eight groups, each with a different testing schedule, and each group counted approximately 335 students.

Mean performance for each group can be seen in **Figure 2*a***. Groups I and II (shown in blue and red, respectively) received their first test on the day of the study session, and they got about 13 correct answers out of a possible 25. Further, they showed little forgetting between Tests 1 and 2, and the same happened between Tests 2 and 3 for Group II, whose members were given a third test 63 days after the initial study session. That is, the relatively flat lines demonstrate stable performance (i.e., very little forgetting) across successive tests spanning the delay period. Group III (shown in green) took the first test a day after the study session, and its performance on the first test was a bit worse than that of Groups I or II; nonetheless, they, too, showed little forgetting between Test 1 and Test 2 (given 2 weeks after the study session). Group IV (shown in purple), which received Test 1 a week after study, did worse than the foregoing groups on Test 1, but it, too, showed no forgetting between Test 1 and Test 2 (given 3 weeks after the study session). Similar observations can be made for Groups V and VI. This pattern of results (since then demonstrated with other participant groups and other materials) has led researchers to suggest that taking a test can at least partially immunize a learner against forgetting. That is, forgetting occurs so that the longer the delay between the study and the first test, the worse the performance on that test. However, on subsequent tests (after Test 1), performance declines over time at a much less pronounced rate than it does in the absence of prior tests.

Another way of looking at the data in **Figure 2** also reveals the role of testing in slowing forgetting. **Figure 2*b*** shows how much the students remembered at a given point in time (e.g., 3 weeks after the study session) as a function of whether prior tests had occurred (and, if so, how many). Spitzer established a pre-experimental group equivalence by another memory measure (not considered here), and all students were treated equally except for their testing schedule. The reader can see that performance on Test 1 showed pronounced declines after a delay, conforming to a standard Ebbinghaus-type forgetting function [Ebbinghaus 1964 (1885)]: That is, in the absence of retrieval practice, pronounced forgetting occurred. The forgetting function for Test 2, however, was shallower. At 21 days after study, the number of correct answers for students taking the first test was lower than that for students who were taking their second test (8.15), which in turn was lower than that for students who were taking their third test (12.18). Taking an initial test markedly affected the subsequent capacity to remember. Further, taking the initial test relatively early—before too much forgetting set in—helped performance more than a delayed initial test.

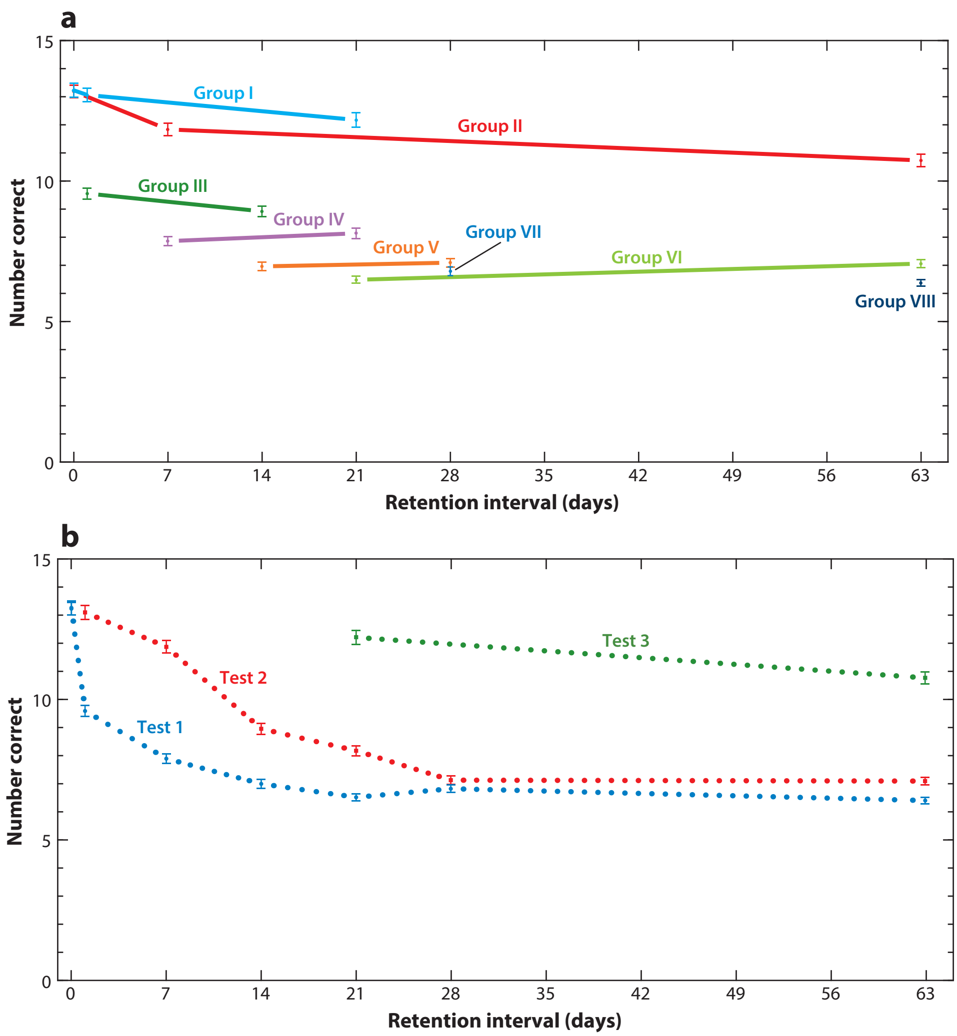

Figure 2

Taking a memory test attenuates forgetting. Error bars represent standard errors of the mean. Data from Spitzer (1939).

These early studies demonstrate the strong effects an initial memory test can exert on what is later remembered. This situation can be viewed as an interesting focus of investigation or as a problematic situation to be avoided in the experimental design stage, depending on the goals of the research. The latter mindset came into favor in the 1940s and following decades (as reviewed in Roediger & Karpicke 2006b). Repeated tests introduced unwanted confounds and were thus considered undesirable. Hence, although the power of retrieval tests was acknowledged in the early twentieth century and before, in the mid-twentieth century there was a considerable temporal gap in research designed to understand this phenomenon, during which time researchers mostly avoided testing people sequentially on the same materials.

In an article entitled "The 'testing' phenomenon: not gone but nearly forgotten," Glover (1989) attempted to redirect attention to this phenomenon as the object of experimental study. He asked

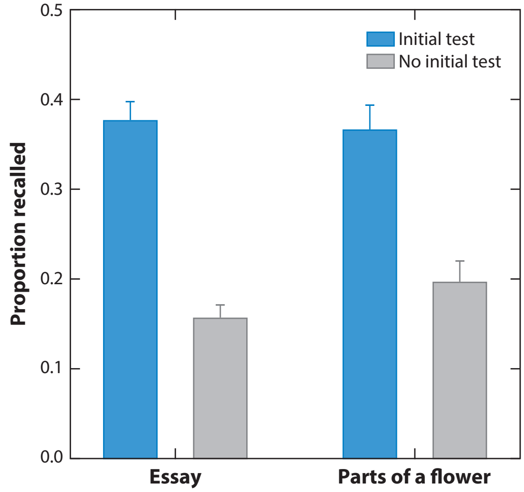

Figure 3

The testing effect. In Glover's (1989) study, final recall was greater when an initial test had occurred. Both college students remembering an essay and seventh-grade students learning parts of a flower benefited from having taken a test (without feedback) 4 days before the final test. Data from Glover (1989).

college students to spend 10 minutes reading a 300-word essay and seventh-grade students to spend 15 minutes studying a diagram of a flower with 12 parts labeled. All students were given a final test 4 days after the study session. In the intervening time, however, about half the students were asked to retrieve the studied material (i.e., freely recalling the essay or labeling the parts of a flower) 2 days after the initial study session, with no instructor feedback given on this initial test. Performance on the final test differed markedly as a function of whether a prior retrieval attempt had occurred (see **Figure 3**). For essays, the students who had taken the initial test outperformed those who had not (0.38 and 0.16 recall probabilities, respectively). Similarly, for the flower diagrams, the students taking an initial test outperformed those who had not (0.37 and 0.20 recall probabilities, respectively). This difference on a final test as a function of whether or not an initial test occurred is referred to as the testing effect.

MODERN INTEREST IN THE POWER OF RETRIEVAL PRACTICE

Modern interest in the testing effect can arguably be traced to two prominent papers published in 2006, both by Henry Roediger and Jeff Karpicke. The first (Roediger & Karpicke 2006a) is an empirical paper reporting an experiment in which college students were asked to read two short science passages. The students studied the first passage for 7 minutes and then studied it again for a second 7-minute segment; for the second passage, participants were asked to study it for 7 minutes and then to spend 7 minutes attempting to remember it (i.e., a free recall test). A final free recall test was given for all passages, but the delay preceding that test varied across participants, from 5 minutes to 2 days or 1 week. As shown in **Figure 4**, the effect of a prior retrieval episode depended upon the delay (i.e., retention interval). Specifically, when the retention interval was only 5 minutes, two study periods produced slightly (but significantly) better recall performance than a study period followed by an initial test (without feedback). After 2 days or 1 week, however,

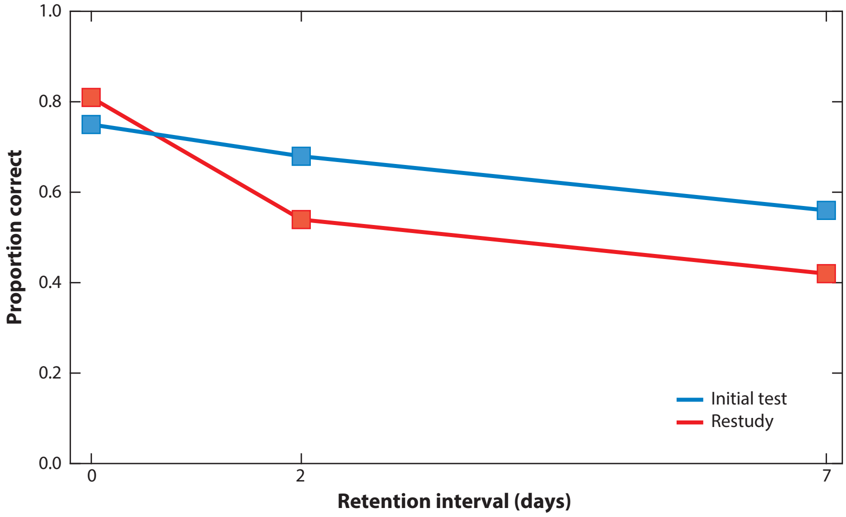

Figure 4

Effects of retrieval practice and retention interval on later recall. When the final test occurred 5 minutes after learning, studying twice proved more beneficial than studying once and then taking an initial test (without feedback). When the final test was delayed 2 days or 1 week after learning, the opposite pattern emerged: Retrieval practice following a single study episode (the initial test condition) was better than two study periods (the restudy condition). Data from Roediger & Karpicke (2006a).

the reverse pattern was observed: Retrieval practice led to recall probabilities that exceeded those in the repeated study condition.

A second experiment confirmed and extended these findings, thereby demonstrating a pronounced, robust role of practicing retrieval in establishing long-term memory. This second experiment manipulated retrieval practice (versus restudy) between participants. When asked to predict (on the first day, after learning) how well they would remember the information on a test one week later, the students who had received an initial test predicted that they would do worse than those who had only studied. This metacognitive confusion might be driven in part by the fact that when the final criterial test is immediate, repeated studying is indeed more beneficial—that is, the students' judgment of their learning was accurate for that moment in time, even though they were asked to make the judgment with respect to their memories one week in the future). Even sophisticated learners can fail to appreciate that retrieval practice is an especially effective learning tool when the test is given after a substantial delay.

The second 2006 paper by these authors was a review in which they defined the testing effect as follows: "Taking a test on material can have a greater positive effect on future retention of that material than spending an equivalent amount of time restudying the material, even when performance on the test is far from perfect and no feedback is given on missed information. This phenomenon of improved performance from taking a test is known as the *testing effect*" (Roediger & Karpicke 2006b, p. 181) (see also Carpenter et al. 2008 for a similar definition). Roediger & Karpicke (2006b) provided an extensive historical overview of the literature and helped focus attention on this relatively neglected finding. A recent retrospective review (Roediger & Karpicke 2018) considers why Roediger & Karpicke's (2006b) article had so much influence. One idea the authors suggest is that pairing the powerful, surprising findings of their first article (Roediger & Karpicke 2006a) with this larger review created a synergistic effect. Another idea is that taking those findings away from a strictly verbal-learning tradition and linking them more to the real-world problems faced by classroom learners (both in the choice of materials and the framing of the studies) helped bring attention to the topic. In the early 2000s, research was bolstered by an influx of funding on applying cognitive psychological principles to educational practice (e.g., by a 10-year, $8.5 million, eight-investigator Collaborative Activity Award by the James S. McDonnell Foundation and by the establishment in 2002 of the US Institute of Education Sciences to encourage evidence-based educational research). In the mid- to late 2000s, interest coalesced around retrieval practice via testing.

TERMINOLOGY

The reader may have noticed that the definition of the testing effect offered by Roediger & Karpicke (2006b) differs from that suggested by Glover (1989) in a critical respect. Both perspectives focus on the benefit to later memory of having taken an initial test, but the comparison condition differs in the two instances. Glover focused on the difference between having taken a test or not (**Figure 3**), whereas Roediger & Karpicke (2006b) conceptualized the difference between having taken an initial test or having restudied the material (**Figure 4**). The latter contrast is arguably more educationally relevant; it also comes closer to equating time-on-task. When the control condition involves restudying, the benefits from practicing retrieval are especially impressive in that restudying usually involves reexperiencing 100% of the material to be learned, whereas tests might refer to only a subset of the material, and usually participants do not successfully retrieve all the information they attempt to retrieve. Therefore, in the absence of feedback, participants in a retrieval practice condition might be considered to be at a disadvantage in terms of sheer exposure.

In past writing (e.g., McDermott et al. 2014b), I have endorsed Glover's definition, which is that "subjects tested between the initial learning episode and the final test . . . outperform subjects only given the final test" (Glover 1989, p. 392). In this review, I do not advocate for one definition over the other, and the following discussion intermixes both types of comparisons. There is inconsistency in the literature about the appropriate control condition—and therefore about the very definition of the term "testing effect" and whether the effect is, or is not, observed. My primary suggestion with respect to this issue is that authors be very clear about how they define the term "testing effect" (if they choose to use it, which is an issue discussed next).

Karpicke has advocated that we move away from the term "testing effect" and focus instead on "retrieval practice" and its effects on later memory (e.g., retrieval-based learning) (see, e.g., Karpicke 2017). The logic behind this argument is that the term testing effect may give inappropriate weight to the need for a formal, external test. Testing suggests the importance of an externally administered assessment, with the learner playing a somewhat passive role as the receiver of a test. Retrieval-based learning puts the emphasis on the learner's cognitive processes, which are (of course) the important ingredient. This term helps underscore that any learner can practice retrieval on their own (with or without an actual test). One can reap the benefits of retrieval practice while walking the dog or doing chores at home; a formal test is not a key part of the equation, nor is any overt response by the learner. That is, covert (silent) retrieval practice is an effective learning tool (Putnam & Roediger 2013).

Another issue, which is especially relevant to the case in which the initial test is followed by a restudy opportunity or correct-answer feedback, concerns the distinction between the direct and indirect (or mediated) effect of retrieval on later retention. Retrieval practice can draw the learner's attention to what they do not know, enhancing the metacognitive understanding of their knowledge. Further, when later study opportunities are provided, prior tests can enable a learner to benefit more from the future re-encoding opportunity (an effect known as test-potentiated learning; see, e.g., Arnold & McDermott 2013a,b) or even to learn more novel information (i.e., test-potentiated new learning; see Chan et al. 2018, Pastötter & Bäuml 2014, Szpunar et al. 2008).

Such effects are important especially in the applied domain, in which the mediated effects of testing can even include such things as incentivizing students to attend class or to ask questions during class (if there are low-stakes daily quizzes, for example). Despite the relevance of this issue, the focus of this review is rather on the direct effect of testing, whereby the retrieval process itself enhances memorability even in the absence of feedback.

GENERALIZABILITY

This section covers controlled laboratory studies of the effect of practicing retrieval on later retrieval accessibility, considering in particular the generalizability of the effect. [For an early review of the recently emerging in-class studies using actual classroom content, the reader is referred to McDaniel et al. (2007).]

The issue of generalizability raises a series of important questions. Does retrieval practice enhance later memory for a variety of materials? Does the effect differ across age groups? Does the format of the test (e.g., short answer or multiple choice) influence the effect? How does retention interval affect the magnitude of the testing effect? Do all people benefit from retrieval practice, or do certain subject characteristics influence whether (or how much) retrieval practice helps later memory? Does retrieval practice enhance learning in actual classroom settings? And does retrieval practice facilitate retrieval only of the practiced information, or might the effect spread to related information? This review is necessarily selective, as the explosion of research in this area precludes a thorough treatment of all these issues, but the interested reader is referred to other recent reviews (e.g., Malmberg et al. 2014) and to several recent meta-analyses (Adesope et al. 2017, Pan & Rickard 2018, Rowland 2014, Schwieren et al. 2017), which have established the robustness of the testing effect and suggested relevant moderators. For example, Rowland (2014) used a random-effects model to compare conditions with testing to control conditions involving an equivalent amount of time restudying, restricting the analysis to laboratory studies because of the greater experimenter control they afford. The overall testing effect was robust (Hedge's $g =$ 0.50).

Materials

The beneficial effects of retrieval practice have been demonstrated across a variety of materials, as reviewed by Dunlosky et al. (2013). The effects on nonsense syllables and biographies have already been discussed above in reference to recitation studies (Gates 1917). Effects of retrieval have also been seen on word lists (Carpenter & DeLosh 2006), paired associates (Allen et al. 1969, Carrier & Pashler 1992), nonverbal materials (Carpenter & Pashler 2007, Kang 2010, Rohrer et al. 2010), and face-name pairings (Morris & Fritz 2000). More educationally relevant stimuli demonstrate similar effects, as seen with prose passages (Glover 1989, Roediger & Karpicke 2006a), science articles from the *National Geographic* (Meyer & Logan 2013), videotaped lectures (Butler & Roediger 2007), and narrated animations (Johnson & Mayer 2009). As discussed below, benefits have also been seen in actual classroom assessments using course materials (e.g., in middle school science and high school history classes).

Does the type of material influence the magnitude of the benefit? Some researchers have attempted to address the issue—for example, Gates (1917) suggested that recitation was more beneficial to the remembering of nonsense syllables than of biographies. In general, the question is difficult to answer due to scaling effects associated with the different materials and the accompanying differences in the outcome measures, as discussed below. Thus far, though, no materials have been identified for which retrieval practice does not help memory (in the absence of floor or ceiling effects or other confounding issues). Rowland's (2014) meta-analysis demonstrates that the benefit of retrieval practice relative to restudy is indeed sensitive to the moderating influence of stimulus type, with prose and paired associates registering greater effects than single words or other materials (for which the number of relevant studies was rather low).

Age Groups

Most of the literature on the testing effect, like most of cognitive psychology, has relied on college students as experimental participants. This effect, though, is also readily found in elementary school children (Fazio & Marsh 2019, Karpicke et al. 2016), in middle school (Roediger et al. 2011a) and high school students (McDermott et al. 2014a), in college classrooms (Lyle & Crawford 2011, Mayer et al. 2009), and in medical students (Larsen et al. 2009). Older adults, too, can benefit from retrieval practice (Meyer & Logan 2013, Rogalski et al. 2014, Tse et al. 2010).

Test Format (and the Importance of Correct-Answer Feedback)

The format of the initial test influences the likelihood of observing a testing effect. When the initial test is free recall, there is more of a cognitive burden on the learner to produce the answer compared to a recognition test, in which the answer is presented and one needs only to identify it as having been studied. Therefore, one might expect initial free recall to be a more powerful learning technique than recognition. The literature on this question is not extensive, but Carpenter & DeLosh (2006) report evidence consistent with the hypothesis that free recall is a more powerful memory enhancer than recognition. In some sense, the outcome of their study is unsurprising: More initial work led to better later performance. On the other hand, though, the success rate of the initial free recall test was lower than the success rate of the recognition test, and no feedback was provided (see Carpenter & DeLosh 2006, experiment 1). To the extent that the testing effect relies upon successful initial retrieval, one might have predicted the opposite pattern (i.e., that recognition would be a more effective memory enhancer). In essence, there are competing forces at play: More retrieval effort is helpful, but if the test is so hard that retrieval attempts are often unsuccessful, such increased effort is less beneficial for later memory (at least in the absence of correct-answer feedback). This issue is discussed by Karpicke (2017; see also Karpicke et al. 2014b), who suggests that the benefit of initial tests (without feedback) can only be seen when the learner is reasonably successful on the initial test. Indeed, Rowland's (2014) meta-analysis demonstrates that the largest testing effect (relative to restudying) is seen when the initial test performance is greater than 75% and no feedback is given; intermediate benefit is seen when the initial test performance is 51–75%; and no reliable advantage appears when the initial test performance is less than 50%. Giving feedback on the initial test boosts its effects on later memorability, overcoming issues with poor immediate performance.

The format of the final test plays a role as well (as can the match or mismatch in format of the initial and final tests, although various factors seem to moderate these effects; see Kang et al. 2007). When the final test is recall, the testing effect is robust; however, when the final test is recognition, sometimes there are benefits of initial recall (Hanawalt & Tarr 1961, Roediger & McDermott 1995) but sometimes not (Darley & Murdock 1971, Lockhart 1975). It appears, though, that even when a testing effect cannot be seen in overall recognition performance, if one breaks down the components of recognition into recollection and familiarity (using methods developed under the dual process framework; see Yonelinas 2002), there is evidence of the testing effect in the recollection component (Chan & McDermott 2007). In terms of overall performance increases, though, recall tests (especially free recall tests) exhibit more reliable, robust benefits (Rowland 2014).

Most learners in applied settings probably think more in terms of multiple-choice, short-answer, and essay tests than of free recall and recognition. The literature here shows clear benefits of initial testing, but which condition is more effective is dependent on whether or not correct-answer feedback occurs at the time of the initial test (and sometimes on the match in format of the initial and final test). When correct-answer feedback occurs at the time of the initial test, there is some evidence that short-answer tests are better, regardless of whether the final test is in the same short-answer format or is a multiple-choice test (Kang et al. 2007). In essence, the feedback helps make up for the lower level of initial performance typically seen in short-answer tests relative to multiple-choice ones, and it allows the beneficial effects of the extra retrieval effort to manifest on the later test.

The section titled Extensions considers a similar question in actual classrooms and examines whether the effects observed in the laboratory hold up under authentic classroom conditions. First, though, I consider a possible negative aspect of multiple-choice tests. Specifically, when a learner encounters a multiple-choice test, they see incorrect answers, which might lead to confusion at a later time (especially if no immediate feedback is given) (Roediger et al. 2011b). After taking an initial multiple-choice test, students tend to produce the incorrect answers they had previously chosen when later given a short-answer test (Marsh et al. 2009, Roediger & Marsh 2005). In a study in which students answered questions from SAT II tests (on biology, chemistry, history), the overall effect of the multiple-choice test was beneficial (i.e., a testing effect occurred); nonetheless, the prior multiple-choice test also enhanced the likelihood of incorrect answers on the final short-answer test (Marsh et al. 2009). In short, the overall effect seems to be positive, but some caution is warranted when multiple-choice tests do not include correct-answer feedback. Students may select the wrong options and learn those incorrect answers.

Retention Interval

The data in **Figure 4** suggest that retention interval is a contributor to whether a testing effect will occur. Specifically, in Roediger & Karpicke's (2006a) experiment, repeated study was more effective than prior testing when the retention interval preceding the final criterial test was short (see also Thompson et al. 1978). Other studies have shown a testing effect even at short interval; for example, Runquist (1983) found a benefit of prior testing after just 10 minutes, although the amount of benefit was much greater after 1 week. Other studies, too, have suggested that the benefit of prior testing increases over time (Toppino & Cohen 2009), and this pattern can also be seen in a meta-analysis (Rowland 2014), which also demonstrates that a retrieval practice effect often does exist at shorter intervals.

How interesting is the influence of retention interval? I suggest that the crossover interaction whereby repeated study shows superior benefits early on but is surpassed by repeated testing after a delay has been one of the aspects of the testing effect that has led investigators to find it a compelling topic of research. This interaction was shown in two separate experiments by Roediger & Karpicke (2006a) (see **Figure 4**), and much has been made of this marked change in the patterns of performance. One might think it could be interesting to try to identify the inflection point or the delay at which the reversal in patterns occurs. Rowland & DeLosh (2015), however, demonstrate that the level of success on the initial retrieval attempt is important in in determining whether an advantage of retrieval practice will appear at short intervals. They suggest that when performance on the initial test is low, there is an unfair advantage for the restudy condition, such that 100% of the items are restudied but only a small fraction of the items are successfully recalled on the initial test. As a result, the confound in number of exposures masks the advantage that is present for the items that are successfully retrieved (see also Karpicke

2017). This logic will become more apparent in the section titled Theories and Frameworks, in which Kornell et al.'s (2011) bifurcation model is discussed. Based on the finding that testing effects is often present at shorter intervals (see Rowland 2014), I agree with Karpicke (2017) that searching for an inflection point will not be informative.

Setting aside the issue of a crossover interaction, we can reconsider whether an ordinal interaction seems to occur such that the benefit of retrieval practice increases over time. The answer here appears to be yes (Rowland 2014). A final observation with respect to retention interval is that the effects of prior testing are long-lasting, as can be seen in **Figures 2** and **4** and as discussed below for even longer intervals in the extensions to classroom settings.

Individual Differences

Researchers have recently considered the question of whether individual differences might influence the magnitude or presence of the testing effect (Brewer & Unsworth 2012, Minear et al. 2018, Pan et al. 2015, Robey 2019). This question derives from an interest in whether retrieval practice can benefit everyone and whether it might be more beneficial for some people than others. Clearly, these are important practical questions, and the fact that retrieval practice is helpful on average across people does not automatically imply that it will benefit everyone (such a conclusion would be an ecological fallacy). Certainly there are individual differences in long-term memory (Unsworth 2019, Zerr et al. 2018), but whether such differences extend to the testing effect remains to be seen.

There are a number of key issues that will need to be addressed before this literature matures. First, does the testing effect exhibit stability (e.g., in test-retest) over time and materials? Do the benefits of retrieval practice behave in a trait-like way? The answer may well be no. One complication is that the testing effect is a difference score between a condition that receives retrieval practice and one that does not, and difference scores are inherently poor in reliability (Draheim et al. 2019). Further, it is often the case that cognitive tasks that produce robust effects in experimental settings are poorly suited for individual differences studies due to low between-participant reliability (Hedge et al. 2018). Even if there are robust positive correlations between the magnitude of one's testing effect across time and materials, can these effects be separated from scaling effects (cf. Pan et al. 2015)? That is, if a given person exhibits a small testing effect but demonstrates 94% recall in the nontested condition, that is not very informative with respect to any meaningful measure, given that the person had very little room for improvement. If the test is more difficult, that same person may benefit greatly from an initial test. For these reasons, attempting to identify individual differences in the testing effect are at this point premature. We first need to know whether stability exists, and if so, whether it exists independently of ceiling and floor effects. This is an important issue, especially in light of recent recommendations that classrooms (and more generally students) adopt retrieval practice as a learning mechanism, as discussed below.

EXTENSIONS

Enhancing Classroom Learning

Some studies have demonstrated testing effects using classroom materials (e.g., Carpenter et al. 2009) or using experimenter-created materials in a simulated classroom setting (Butler & Roediger 2007, Duchastel & Nungester 1982); to date, however, very few studies have attempted to address whether introduction of in-class quizzing can improve actual course grades.

Can low-stakes testing (quizzing) be incorporated into normal classroom procedures to enhance student learning? The laboratory studies reviewed above are suggestive, but they are not sufficient evidence for firm recommendations. One big difference is that most of the laboratory studies present information once and then quiz (or review) once. In the classroom, however, integrated materials are repeatedly encountered in teachers' lectures and students' homework. Further, it is customary to provide correct-answer feedback, which makes interpretation of the results a little murkier: It is difficult to know whether any potential effects arise from the test itself or from the feedback and its influence on future encounters with the information. A "clean" testing effect experiment, whereby the participant studies (tests or not) and then (with no intervening encounters) takes a final test, would be difficult in a classroom, but it would also lose its educational relevance. Hence, the work below is inspired by an understanding of retrieval practice effects but necessarily loosens some of the controls at work in laboratory studies.

McDermott et al. (2014a) tested students in a Midwestern suburban public middle school and its affiliated high school. In a seventh-grade science class, students were given quizzes in multiple-choice and short-answer formats such that the format type was intermixed throughout the quizzes. Quizzes were administered by a research assistant while the teacher took a short break. Importantly, items were counterbalanced across the various classroom sections so that a given item would be quizzed in short-answer format for some students, multiple-choice format for others, and not at all for other students. Because the research assistant controlled the quiz construction and administration, the teacher did not know when teaching a given section which items would be assigned to which conditions (to control for any instructional bias).

Consider first an experiment looking at whether the format of the low-stakes quizzes influenced the scores on the in-class unit exam. Specifically, across six sections of seventh-grade science, the students took three quizzes prior to the unit exam; all quizzes occurred after the relevant instruction, and the final quiz occurred as a review a day before the unit exam. The question format on the unit exam also varied. The question stem was identical for multiple-choice and short-answer tests (e.g., "What are two characteristics of angiosperms?"), but in one case the students supplied the answer themselves (short-answer format) and in another they selected from among a set of possible alternatives (multiple-choice format).

Figure 5*a* shows that the students benefited from the quizzes. Specifically, for the questions in multiple-choice format on the unit exam, the likelihood of a correct answer was greater for either quiz type (0.84 and 0.85 proportion of correct answers for multiple choice and short answer, respectively) compared to when an initial quiz had not occurred (0.69 correct). Similarly, when the unit exam question was in the short-answer format, students performed better if the question had been previously quizzed (0.60 and 0.61 for multiple choice and short answer, respectively) relative to not quizzed (0.40 correct). Importantly, the type of quiz did not influence the magnitude of the benefit; what was important was that a quiz with immediate correct-answer feedback had occurred.

To establish whether this effect generalized to other classrooms, we conducted an analogous experiment in a high school US history class. The design was very similar, except that students took only two initial quizzes, and the question was changed between the first quiz, the second one, and the unit exam. For example, a target fact was queried in the following ways: "What act of congress prevented discrimination in employment and public accommodations, and provided the federal government with the powers to enforce desegregation?" or "Because it involves discrimination in employment, refusing to hire someone solely because they are African American would be a violation of which act of congress?" or "Because it involves discrimination in public accommodations, requiring African American people to use a separate water fountain at a state park would be a violation of which act of congress?" Each student's quiz intermixed multiple-choice and short-answer questions, and the question stems were identical for the two formats.

Figure 5*b* shows that again students benefited from the quizzes. Further, the quiz format did not influence the amount of benefit; what mattered was that the facts had been quizzed. These two

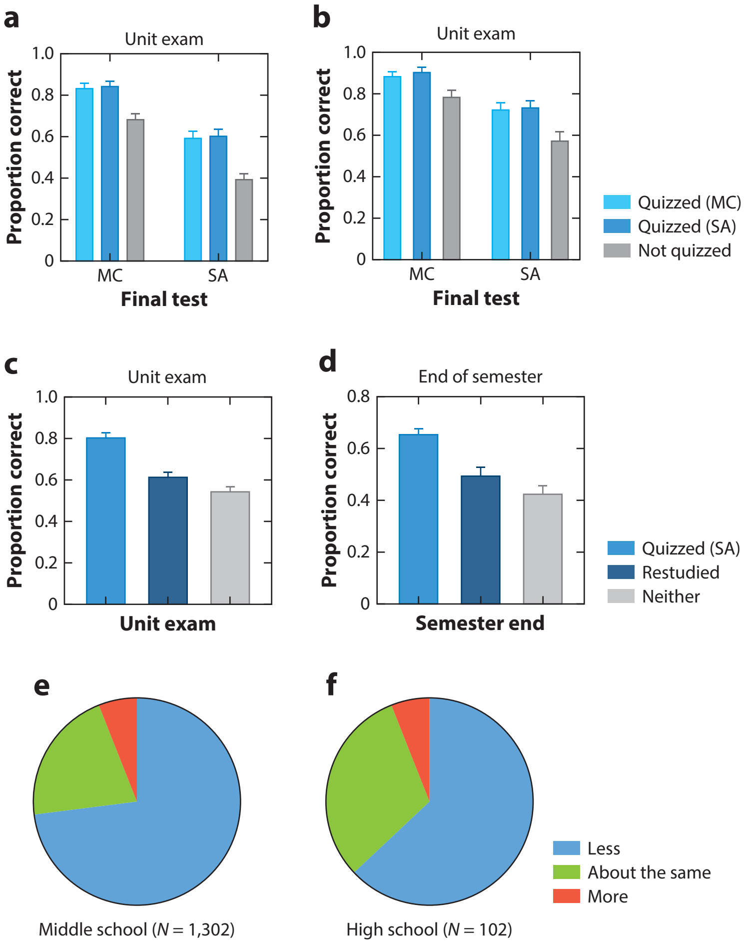

Figure 5

Retrieval practice effects in the classroom. In true experiments performed in middle school classrooms, performance on the unit exam and end-of-semester exam was better for items that had been quizzed than for items that had not appeared on a quiz (panels *a–d*). The specific quiz format (multiple choice or short answer) was not critical in determining a retrieval practice benefit. Further, students performed better for items that had been quizzed than for items that had been restudied in the form of a quiz answer key; the effort to retrieve is important. Finally, frequent classroom quizzing reduced test anxiety for most middle school and high school students (panels *e*, *f*). Abbreviations: MC, multiple choice; SA, short answer. Data from McDermott et al. (2014a).

experiments show that quizzes in a classroom can influence grades on a unit exam. Regardless of type of question on unit exam, prior quizzing (in either multiple-choice or short-answer format) benefited later performance on a graded in-class test.

One concern might be that the quizzes simply directed the students' attention to relevant facts that would be tested on the unit exam. Would simply reencountering the information give the same benefit (but in the absence of retrieval practice)? To address this possibility, before taking a unit exam (in the short-answer format), seventh-grade science students were alternatively given short-answer quizzes, asked to restudy the same facts (i.e., receiving the question stem accompanied by an ideal answer), or neither of the two. Again, all students took the quizzes, and the items were rotated across conditions across student sections. As can be seen in **Figure 5*c***, restudying the selected facts did not provide the same enhancement on the unit exam (see also Roediger et al. 2011a); taking an initial quiz led to much better performance than doing what amounts to studying the answer key of a quiz (0.81 and 0.62 proportions correct, respectively), with correct performance in the no-test condition at 0.55.

Finally, we can ask whether these benefits are long-lasting. When a surprise end-of-semester test was given to these same students, the earlier benefits of quizzing remained (**Figure 5*d***). Taking an initial quiz facilitated performance relative to restudying (0.66 and 0.50, respectively).

Across these studies we can see clear benefits of retrieval practice in a true classroom environment, where learning unfolds over time and involves repeated exposure to the materials. The specific details of the quizzing situation are not as important as the fact that quizzing occurred. The effect was seen in middle school and high school, in science and history, after multiple-choice and short-answer quizzes, on unit exams (in multiple-choice and short-answer formats) and on an unexpected end-of-semester exam. Question wording can change from quiz to quiz to ensure that students are learning actual content beyond any specific wording. Providing an answer key (but not asking students to attempt to retrieve) is not as effective. In other studies, we have also repeatedly demonstrated that core classroom content can be better remembered on unit exams after low-stakes multiple-choice quizzing with immediate correct-answer feedback (McDaniel et al. 2007, 2011, 2013; Roediger et al. 2011a).

The inclusion of feedback in the aforementioned studies is worthy of consideration. If the goal had been to examine the effect of retrieval practice alone (without any potential contaminating effects), it would have been desirable to not include feedback in these studies. However, the goal of these studies was to enhance student learning, and the administration of quizzes without feedback in middle school may have proven difficult as a first step. Nonetheless, caution is warranted regarding the exact mechanism of improvement: The benefit might be the direct effect of the quizzes themselves, or it might be that the quizzes allowed the students to gain more from subsequent encounters with the materials (in later lectures, when studying at home). Regardless of the mechanism, the studies demonstrate real-world benefits of retrieval practice for learning.

A common concern, especially for children, is that frequent quizzing might enhance student anxiety about grades. Agarwal et al. (2014) addressed this question by asking students in Midwestern middle and high schools whether the in-class low-stakes quizzes had made them more nervous for their unit tests. As can be seen in **Figure 5*e*,*f***, most middle school ($N = 1{,}302$) and high school ($N = 102$) students agreed: Quizzing did not make them any more nervous for the unit test. Indeed, most students reported decreased test anxiety after having worked with the materials on the quizzes.

Incidentally, I have been administering daily end-of-class quizzes in an undergraduate Human Learning and Memory class for 19 years, and despite common early-semester misgivings about the quizzes, a frequent comment on end-of-semester evaluations is that students appreciate the role the low-stakes quizzes played in helping them learn. Specifically, the quizzes incentivize class attendance and the asking of questions (the indirect or mediated effect of testing) in addition to the powerful direct effects. A suggestion for instructors who choose to give quizzes is to include a lecture (or a section of a lecture) showing how powerful quizzes can be for enhancing learning.

Similar conclusions were reached in a sample of college students taking an online course in Brain & Behavior (Thomas et al. 2018). When quizzing (with feedback) was incorporated, exam performance was enhanced for that material (relative to the nonquizzed materials), and this benefit accrued even when the type of question changed from the quiz to the exam (cf. Foss & Pirozzolo 2017). That is, exam questions that probed applications of the concept benefited from prior quizzing of relevant facts; alternately, exam questions focusing on facts improved after students were quizzed on the applications of those facts. Similar benefits of in-class quizzes in college classes have been seen in a Psychology Research Methods class (Bjork et al. 2014), in an Introductory Psychology class (Batsell et al. 2017), and in a Statistics for Psychology class (Lyle & Crawford 2011), all of them at large state universities. Bjork et al. (2014) demonstrated that the benefits extended not just to the actual information tested but also to related information. This phenomenon is considered below.

Retrieval-Induced Facilitation: Enhancing Memory for Related, Nontested Information

One concern about the use of quizzes in classroom settings is that one is enhancing the memorability of selected information at the expense of other important information. But might retrieval practice benefits spread to related, nontested information? The answer is—in some circumstances—yes.

Chan et al. (2006) asked university students to read an article much like what would appear in a college textbook, and they were told to expect a memory test. Then they answered 22 quiz questions (short-answer format) without feedback two times (i.e., did two quizzes with the same items) or reread the facts targeted by the quizzes. A day later, retrieval for target facts related to the tested facts, but not tested themselves, was 9% greater than retrieval for nontested facts that were unrelated to the tested facts (59% and 50% recall, respectively). Simply reading the target facts did not produce facilitation, however (49% recall). Chan et al. (2006) provided follow-up evidence that the initial quizzes led students to retrieve not only the specific answer to the question but also related information, which then received a benefit.

The phenomenon of retrieval-induced facilitation stands in obvious contrast to another phenomenon known as retrieval-induced forgetting (Anderson et al. 1994). In the latter situation, practicing retrieval of a subset of studied information can harm performance for the nontested information. Obviously, if retrieval-induced forgetting were to occur in a classroom, that would be problematic. Chan (2009) set out to understand when one might expect to see retrieval-induced forgetting and when one might expect facilitation. Specifically, Chan noted that retrieval-induced forgetting tends to be short-lived (MacLeod & Macrae 2001) and can be eliminated if participants are asked to integrate the materials (Anderson & McCulloch 1999). In a prior study in which Chan et al. (2006) had observed retrieval-induced facilitation, a relatively longer one-day delay had been used. Therefore, Chan (2009) hypothesized that if retention intervals were manipulated, one might observe retrieval-induced forgetting at a short delay (20 minutes) when integration was discouraged but retrieval-induced facilitation at a longer delay (24 hours) when integration was encouraged. Indeed, that pattern was observed. The finding of retrieval-induced facilitation at long retention intervals was replicated using textbook materials that encourage integration (Chan 2010) (see **Figure 6**). After 20 minutes, no retrieval-induced facilitation and no testing effect was observed; at longer delays both emerged robustly (see also Cranney et al. 2009).

The phenomenon of retrieval-induced facilitation has ties to the literature on transfer, which has shown that retrieval practice can enhance near transfer (Butler 2010, Kang et al. 2011, Rohrer et al. 2010). Such benefits were shown compellingly by Karpicke & Blunt (2011), who compared how retrieval practice, restudy, and concept mapping (i.e., generating diagrams to depict relations among concepts) would influence later performance for the specific concepts in a science text and for related questions requiring an inference. On a final test administered a week after the initial experience, when given verbatim or inference questions, university students showed the greatest benefits from retrieval practice, followed by concept mapping and repeated study (which did not differ). A second experiment demonstrated that even if the final criterial test was to create

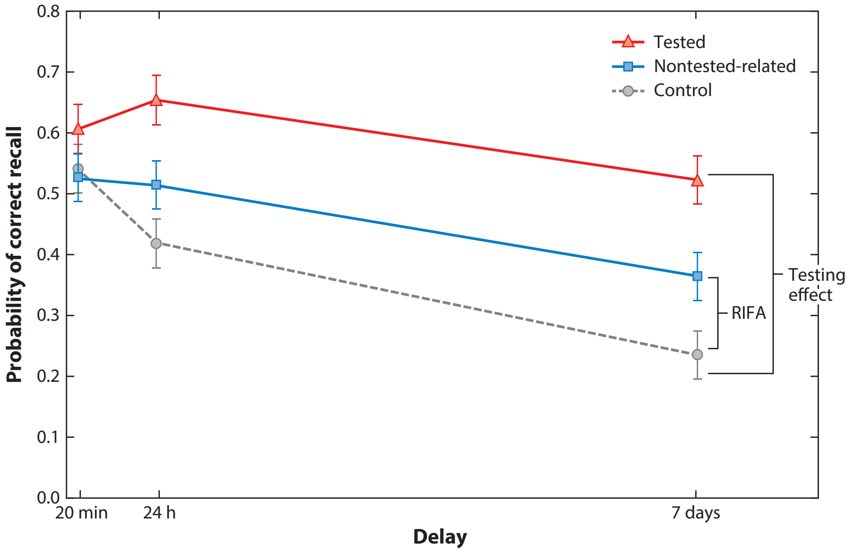

Figure 6

Retrieval-induced facilitation (RIFA) is long-lasting. Figure adapted with permission from Chan (2010).

a concept map (with no external cues provided), having done retrieval practice earlier led to better performance than having created a concept map. The authors point out that concept mapping could be used in combination with retrieval practice, and they have later shown the effectiveness of this combination in elementary school learning (Karpicke et al. 2014a).

THEORIES AND FRAMEWORKS

One seemingly straightforward framework that has been applied to understanding retrieval practice effects is the transfer-appropriate processing framework (Morris et al. 1977). The basic claim is that learning activities are neither inherently good nor bad but need to be understood in the context of the eventual retrieval environment. For example, deep or meaningful processing is usually a more effective encoding strategy than shallow processing (e.g., focusing on the meaning of encoded words is typically more effective than focusing on the phonology), but the transfer-appropriate processing framework captures the finding that if the final criterial test requires memory for the sound of words, having encoded words with respect to phonology is a superior encoding strategy.

When this framework is applied to the domain of retrieval practice, the overarching idea is that practicing retrieval is helpful because it aligns with the eventual goal of facilitating later retrieving. More specifically, a match between the type of retrieval practice (e.g., answering questions on a short-answer test or a multiple-choice test) might be expected to be an important predictor of the magnitude of the retrieval practice effect. As seen above, however, a simple match in test type is not always the most desirable situation. Rather, at times a short-answer test can better prepare a student for a later multiple-choice test. As such, this framework is intuitively appealing but fails to capture the whole of the relevant findings.

Another conceptualization is that retrieval effort strengthens memory, and the more retrieval effort one puts out, the better the preparation for later retrieval (Bjork 1975). As such, retrieval effort is considered a "desirable difficulty" (Bjork 1994). This framework, like transfer-appropriate processing, is intuitively appealing and captures some of the basic findings in the literature (e.g., Roediger & Karpicke 2006a). Perhaps for this reason, it is one of the more popular frameworks for discussing the benefits of retrieval practice. A major challenge, though, is that the level of retrieval effort cannot be defined independent of its eventual effect on retrievability (cf. Pyc & Rawson 2009); likewise, it is not clear ahead of time which difficulties are desirable and which are undesirable and whether the answer to that question differs from person to person.

A third idea—the elaborative retrieval hypothesis—is that practicing retrieval tends to involve activation of related concepts, as discussed above. This activation of related concepts enhances the likelihood of future recall by adding to the number of possible retrieval routes at that later recall time (Carpenter 2009). The idea is that when attempting to retrieve (say) "carrot" in response to the cue "vegetable," a person may search memory for plausible answers ("cucumber," "lettuce," "kale") before arriving at the correct answer. Then on the final criterial test, the learner may benefit from those prior ineffective retrieval attempts, which can help lead to the final answer. One mystery in this logic is that the principle of cue overload (Watkins 1975) suggests that these plausible answers may later interfere with the final retrieval attempt instead of helping it, and it is not entirely clear how these possible answers will facilitate recall. The theory also appears most relevant to verbal materials, and it is not clear how well it would explain nonverbal retrieval practice. The effect sizes giving rise to this hypothesis were not very robust (see Carpenter 2009, figure 1), and attempts to directly test the hypothesis have not revealed support for the generality of the framework (e.g., Karpicke & Blunt 2011).

A fourth proposal—the episodic context account—draws on the idea that retrieval involves reinstatement of the learning context (Karpicke et al. 2014b). This theory assumes (as do many models of memory) that temporal context is a relevant feature encoded into memory, and this context changes slowly over time. Further, when attempting to retrieve, people attempt to reinstate the former temporal context to help the retrieval search process. Thus, on an initial test, when people remember an item from an earlier context (context A), that memory representation becomes combined with the existing context (context B) to form a composite. Finally, when trying to remember again at a later time, somehow the composite context representation allows a restriction of the memory search set, which is beneficial to retrieval. This account is appealing in that it offers a mechanism, although it shares many of the difficulties of the prior approaches in that independently defining its core constructs (e.g., context) is difficult or impossible. Further, it is not entirely clear how combining contexts would simplify or restrict a search set compared to retrieving a single prior context.

A final framework for thinking about the testing effect is the bifurcation model (Kornell et al. 2011). The idea begins with a conceptualization of item distributions. After a person studies a set of items, there will be variability in the resulting memory strength of those items (see **Figure 7*a*,*b***). For participants who then restudy the items, the resulting memory strengths will tend to increase such that the distribution of memory strengths will be shifted to the right (**Figure 7*c***). If instead participants are tested on those items, then some of the items will benefit quite a lot from enhanced memory strengths (i.e., those that are correctly retrieved); the other items will receive no benefit from the (failed) retrieval attempt (**Figure 7*d***). That is, the distribution will become bifurcated. As time passes before the final criterial test, all items will have reduced memory strength due to forgetting; the greater the passage of time, the more shift in the distribution due to forgetting.

A key observation that this conceptualization can accommodate is the finding that the benefits of retrieval practice are sometimes not apparent at short delays but are readily observed after longer delays. This can be seen in **Figure 7*e–h***. After a short delay, 70% of the restudied items have memory strengths that surpass the threshold for recallability, whereas after a longer delay only 43% of the items have enough memory strength to be recalled (**Figure 7*e*,*g***). This can be

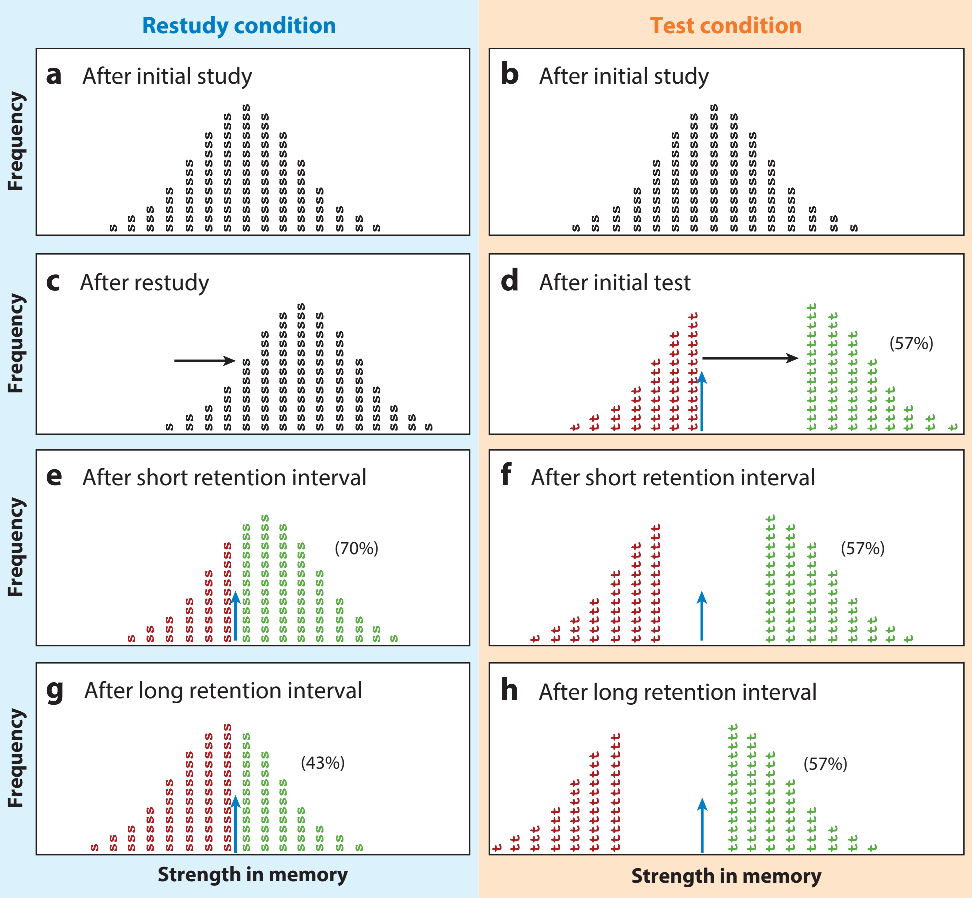

Figure 7

The bifurcation model of forgetting for items previously restudied (*left column*) or previously tested (*right column*). Figure adapted with permission from Kornell et al. (2011).

contrasted with the situation in which a test has taken place instead of restudy (**Figure 7*f*,*h***). Those items that were retrieved on the initial test exhibit high memory strengths that surpass the recallability threshold (despite some forgetting taking place relative to the initial test period). Further, the benefit obtained from that initial test is robust enough that all 57% of the items that were initially tested are above the recall threshold at both the short and long delays. This conceptualization can accommodate the finding that initial tests can appear to immunize against forgetting (e.g., see **Figure 2*a***). Further, this framework also accommodates the phenomenon whereby restudy shows superior benefits compared to testing after a short retention interval but the reverse is true after a longer delay (see, e.g., **Figure 4**).

Space precludes a thorough review of these theoretical frameworks, but the interested reader is referred to several reviews for more extensive coverage (Karpicke 2017, Karpicke et al. 2014b, Nunes & Karpicke 2015). At present, none of the existing theories or frameworks can capture the effects in the entire literature, and most often the conceptual frameworks do not make clear predictions. Future work will need to flesh out a priori predictions with the eventual goal of more precise understanding of how and why retrieval practice aids memory. One challenge is the difficulty in operationally defining key aspects of the relevant features: what processing is transfer appropriate, what difficulties are desirable, what retrieval routes are and how they proliferate, how to pre-experimentally define context, and how context reinstatement changes retrievability. Formal theories and clearly testable predictions remain a goal for the future.

Although outside the scope of this review, it is worth noting that there is a small literature focusing on the neural substrates of the testing effect (e.g., van den Broek et al. 2013, Wing et al. 2013; for a review, see van den Broek et al. 2016) and adjacent phenomena (e.g., test-potentiated learning; see Nelson et al. 2013, Vestergren & Nyberg 2014). One complicating factor is that there is no consensus on what the key contrasts of interest should be, and the specific operationalization of terms differs widely across investigations. For example, some studies have compared the neural activity during initial tests to the activity during restudy trials in a variation of a subsequent memory design (Wagner et al. 1998). Other approaches have examined how neural activity changes across repeated retrieval practice opportunities (e.g., Hashimoto et al. 2011). When one considers the varieties of paradigms that can be used to investigate the power of retrieval practice for later memory, the multiple definitions of the testing effect, and the complexities considered in this section, it is perhaps unsurprising that to date neuroimaging studies have not greatly enhanced our theoretical understanding of the mechanisms underlying the testing effect.

REMAINING PUZZLES: FUTURE DIRECTIONS

The literature on how and when retrieval practice facilitates later retrieval has matured markedly over the last 15 years. There are, however, significant gaps in our understanding. As reviewed above, theoretical accounts of the phenomenon are lacking in precision and predictive power. Further, it is unclear whether retrieval practice offers similar benefits for all learners or some learners will be especially sensitive to the manipulation. Perhaps some of them will show no benefit at all.

The issue of how to balance retrieval effort with retrieval success is nontrivial. For a learner to benefit from retrieval practice, the goal is to make the retrieval effort difficult, but not so difficult that the retrieval attempt fails. Titrating questions to hit this sweet spot is a challenging issue, and even when that goal is accomplished on average (i.e., for most items for average learners), it does not necessarily apply well to learners toward the top or bottom of the distribution for whom those items might be too easy or too difficult, respectively.

Also lacking is a better understanding of the mechanisms behind some of the basic effects considered in this review. It has been argued that testing forestalls forgetting, unlike many other independent variables (e.g., meaningfulness) that do not influence forgetting rate (Underwood 1972, Wixted 2004). That testing does so is interesting, but why it does so is a mystery. Similarly, why does retrieval-induced facilitation occur? To what extent does conscious search play a role, or might spreading activation cause it?

With respect to harnessing the power of retrieval practice to enhance education, perhaps the biggest hurdle is to get learners to realize the phenomenon exists (see the sidebar titled Implementation in the Real World). Whether for an older adult attempting to learn dietary guidelines associated with a newly diagnosed health condition or for a sixth-grade student struggling to get through social studies, an appreciation that retrieval practice aids the learning process is a first step in getting people to incorporate that approach as a study strategy. Current educational practice does not typically include explicit instruction on how to learn, and incorporating learning strategies into elementary school classrooms might bring large benefits down the road. A key point to highlight is that the retrieval practice does not need to be in the form of an externally administered test; simple, covert retrieval practice is sufficient for solidifying learning.

IMPLEMENTATION IN THE REAL WORLD

Resources exist for the layperson interested in these effects. For example, the Retrieval Practice website (**https://www.retrievalpractice.org**) contains a free retrieval practice guide, in addition to other evidence-based learning strategies and explanations. Popular books (e.g., Agarwal & Bain 2019, Brown et al. 2014) have had some success at enhancing awareness. Incorporation of daily quizzing by instructors in the classroom (especially if accompanied by an explanation of the purpose of the quizzes) can prove to students that retrieval practice—however informal—can be a powerful technique in solidifying learning.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Preparation of this article was supported by a Collaborative Activity Award from the James S. McDonnell Foundation. I am grateful to Nate Anderson, Jason Chan, Sean Kang, Roddy Roediger, Oyku Uner, and Chris Zerr for comments on an earlier version of the manuscript.

LITERATURE CITED

Abbott EE. 1909. On the analysis of the factor of recall in the learning process. *Psychol. Monogr.* 11:159–77

Adesope OO, Trevisan DA, Sundararajan N. 2017. Rethinking the use of tests: a meta-analysis of practice testing. *Rev. Educ. Res.* 87(3):659–701

Agarwal PK, Bain P. 2019. *Powerful Teaching: Unleash the Science of Learning*. San Francisco, CA: Jossey-Bass

Agarwal PK, D'Antonio L, Roediger HL III, McDermott KB, McDaniel MA. 2014. Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *J. Appl. Res. Mem. Cogn.* 3(3):131–39

Allen GA, Mahler WA, Estes WK. 1969. Effects of recall tests on long-term retention of paired associates. *J. Verbal Learn. Verbal Behav.* 8:463–70

Anderson MC, Bjork RA, Bjork EL. 1994. Remembering can cause forgetting: retrieval dynamics in long-term memory. *J. Exp. Psychol. Learn. Mem. Cogn.* 20(5):1063–87

Anderson MC, McCulloch KC. 1999. Integration as a general boundary condition on retrieval-induced forgetting. *J. Exp. Psychol. Learn. Mem. Cogn.* 25:608–29

Arnold KM, McDermott KB. 2013a. Free recall enhances subsequent learning. *Psychon. Bull. Rev.* 20(3):507–13

Arnold KM, McDermott KB. 2013b. Test-potentiated learning: distinguishing between direct and indirect effects of tests. *J. Exp. Psychol. Learn. Mem. Cogn.* 39(3):940–45

Bacon F. 2000 (1620). *Novum Organum*, transl. LJM Silverthorne. Cambridge, UK: Cambridge Univ. Press

Batsell WR, Perry JL, Hanley E, Hostetter AB. 2017. Ecological validity of the testing effect: the use of daily quizzes in introductory psychology. *Teach. Psychol.* 44(1):18–23

Bjork EL, Little JL, Storm BC. 2014. Multiple-choice testing as a desirable difficulty in the classroom. *J. Appl. Res. Mem. Cogn.* 3(3):165–70

Bjork RA. 1975. Retrieval as a memory modifier: an interpretation of negative recency and related phenomena. In *Information Processing and Cognition: The Loyola Symposium*, ed. RL Solso, pp. 123–44. Mahwah, NJ: Lawrence Erlbaum Assoc.

Bjork RA. 1994. Memory and metamemory considerations in the training of human beings. In *Metacognition: Knowing About Knowing*, ed. J Metcalfe, AP Shimamura, pp. 185–205. Cambridge, MA: MIT Press

Blasiman RN, Dunlosky J, Rawson KA. 2017. The what, how much, and when of study strategies: comparing intended versus actual study behaviour. *Memory* 25(6):784–92

Brewer GA, Unsworth N. 2012. Individual differences in the effects of retrieval from long-term memory. *J. Mem. Lang.* 66(3):407–15

Brown PC, McDaniel MA, Roediger HL III. 2014. *Make it Stick: The Science of Successful Learning*. Cambridge, MA: Belknap Press

Butler AC. 2010. Repeated testing produces superior transfer of learning relative to repeated studying. *J. Exp. Psychol. Learn. Mem. Cogn.* 36(5):1118–33

Butler AC, Roediger HL III. 2007. Testing improves long-term retention in a simulated classroom setting. *Eur. J. Cogn. Psychol.* 19(4–5):514–27

Carpenter SK. 2009. Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *J. Exp. Psychol. Learn. Mem. Cogn.* 35(6):1563–69

Carpenter SK, DeLosh EL. 2006. Impoverished cue support enhances subsequent retention: support for the elaborative retrieval explanation of the testing effect. *Mem. Cogn.* 34(2):268–76

Carpenter SK, Pashler H. 2007. Testing beyond words: using tests to enhance visuospatial map learning. *Psychon. Bull. Rev.* 14(3):474–78

Carpenter SK, Pashler H, Cepeda NJ. 2009. Using tests to enhance 8th grade students' retention of U.S. history facts. *Appl. Cogn. Psychol.* 23(6):760–71

Carpenter SK, Pashler H, Wixted JT, Vul E. 2008. The effects of tests on learning and forgetting. *Mem. Cogn.* 36(2):438–48

Carrier M, Pashler H. 1992. The influence of retrieval on retention. *Mem. Cogn.* 20(6):633–42

Chan JCK. 2009. When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *J. Mem. Lang.* 61(2):153–70

Chan JCK. 2010. Long-term effects of testing on the recall of nontested materials. *Memory* 18(1):49–57

Chan JCK, McDermott KB. 2007. The testing effect in recognition memory: a dual process account. *J. Exp. Psychol. Learn. Mem. Cogn.* 33(2):431–37

Chan JCK, McDermott KB, Roediger HL III. 2006. Retrieval-induced facilitation: initially nontested material can benefit from prior testing of related material. *J. Exp. Psychol. Gen.* 135(4):553–71

Chan JCK, Meissner CA, Davis SD. 2018. Retrieval potentiates new learning: a theoretical and meta-analytic review. *Psychol. Bull.* 144(11):1111–46

Cranney J, Ahn M, McKinnon R, Morris S, Watts K. 2009. The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *Eur. J. Cogn. Psychol.* 21(6):919–40

Darley CF, Murdock BB. 1971. Effects of prior free recall testing on final recall and recognition. *J. Exp. Psychol. Learn.* 91(1):66–73

Draheim C, Mashburn CA, Martin JD, Engle RW. 2019. Reaction time in differential and developmental research: a review and commentary on the problems and alternatives. *Psychol. Bull.* 145(5):508–35

Duchastel PC, Nungester RJ. 1982. Testing effects measured with alternate test forms. *J. Educ. Res.* 75(5):309–13

Dunlosky J, Rawson KA, Marsh EJ, Nathan MJ, Willingham DT. 2013. Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychol. Sci. Public Interest* 14(1):4–58

Ebbinghaus H. 1964 (1885). *Memory: A Contribution to Experimental Psychology*, transl. HA Ruger, CE Bussenius, ER Hilgard. New York: Dover

Fazio LK, Marsh EJ. 2019. Retrieval-based learning in children. *Curr. Dir. Psychol. Sci.* 28(2):111–16

Foss DJ, Pirozzolo JW. 2017. Four semesters investigating frequency of testing, the testing effect, and transfer of training. *J. Educ. Psychol.* 109(8):1067–83

Gardiner JM. 2007. Retrieval: on its essence and related concepts. In *Science of Memory: Concepts*, ed. HL Roediger III, Y Dudai, SM Fitzpatrick, pp. 221–24. Oxford, UK: OPxford Univ. Press

Gates AI. 1917. *Recitation as a Factor in Memorizing*. New York: Science Press

Glover JA. 1989. The "testing" phenomenon: not gone but nearly forgotten. *J. Educ. Psychol.* 81(3):392–99

Hanawalt NG, Tarr AG. 1961. The effect of recall upon recognition. *J. Exp. Psychol.* 62(4):361–67

Hashimoto T, Usui N, Taira M, Kojima S. 2011. Neural enhancement and attenuation induced by repetitive recall. *Neurobiol. Learn. Mem.* 96:143–49

Hedge C, Powell G, Sumner P. 2018. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* 50:1166–86

James WJ. 1890. *The Principles of Psychology*. New York: Dover
Johnson CI, Mayer RE. 2009. A testing effect with multimedia learning. *J. Educ. Psychol.* 101(3):621–29
Kang SHK. 2010. Enhancing visuospatial learning: the benefit of retrieval practice. *Mem. Cogn.* 38(8):1009–17
Kang SHK, McDaniel MA, Pashler H. 2011. Effects of testing on learning of functions. *Psychon. Bull. Rev.* 18(5):998–1005
Kang SHK, McDermott KB, Roediger HL III. 2007. Test format and corrective feedback modify the effect of testing on long-term retention. *Eur. J. Cogn. Psychol.* 19(4–5):528–58
Karpicke JD. 2017. Retrieval-based learning: a decade of progress. In *Learning and Memory: A Comprehensive Reference*, ed. JH Byrne, pp. 487–514. Amsterdam: Elsevier
Karpicke JD, Blunt JR. 2011. Retrieval practice produces more learning than elaborative studying with concept mapping. *Science* 331(6018):772–75
Karpicke JD, Blunt JR, Smith MA. 2016. Retrieval-based learning: positive effects of retrieval practice in elementary school children. *Front. Psychol.* 7:350
Karpicke JD, Blunt JR, Smith MA, Karpicke SS. 2014a. Retrieval-based learning: the need for guided retrieval in elementary school children. *J. Appl. Res. Mem. Cogn.* 3(3):198–206
Karpicke JD, Butler AC, Roediger HL III. 2009. Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory* 17(4):471–79
Karpicke JD, Lehman M, Aue WR. 2014b. Retrieval-based learning: an episodic context account. In *Psychology of Learning and Motivation*, Vol. 61, ed. BH Ross, pp. 237–84. Amsterdam: Elsevier
Köhler W. 1947. *Gestalt Psychology*. New York: Liveright
Kornell N, Bjork RA. 2007. The promise and perils of self-regulated study. *Psychon. Bull. Rev.* 14(2):219–24
Kornell N, Bjork RA, Garcia MA. 2011. Why tests appear to prevent forgetting: a distribution-based bifurcation model. *J. Mem. Lang.* 65(2):85–97
Larsen DP, Butler AC, Roediger HL III. 2009. Repeated testing improves long-term retention relative to repeated study: a randomised controlled trial. *Med. Educ.* 43(12):1174–81
Lockhart RS. 1975. The facilitation of recognition by recall. *J. Verbal Learn. Verbal Behav.* 14:253–58
Lyle KB, Crawford NA. 2011. Retrieving essential material at the end of lectures improves performance on statistics exams. *Teach. Psychol.* 38(2):94–97
MacLeod MD, Macrae CN. 2001. Gone but not forgotten: the transient nature of retrieval-induced forgetting. *Psychol. Sci.* 12:148–52
Malmberg KJ, Lehman M, Annis J, Criss AH, Shiffrin RM. 2014. Consequences of testing memory. In *Psychology of Learning and Motivation*, Vol. 61, ed. BH Ross, pp. 285–313. Amsterdam: Elsevier
Marsh EJ, Agarwal PK, Roediger HL III. 2009. Memorial consequences of answering SAT II questions. *J. Exp. Psychol. Appl.* 15(1):1–11
Mayer RE, Stull A, DeLeeuw K, Almeroth K, Bimber B, et al. 2009. Clickers in college classrooms: fostering learning with questioning methods in large lecture classes. *Contemp. Educ. Psychol.* 34(1):51–57
McCabe J. 2011. Metacognitive awareness of learning strategies in undergraduates. *Mem. Cogn.* 39(3):462–76
McDaniel MA, Agarwal PK, Huelser BJ, McDermott KB, Roediger HL III. 2011. Test-enhanced learning in a middle school science classroom: the effects of quiz frequency and placement. *J. Educ. Psychol.* 103(2):399–414
McDaniel MA, Roediger HL III, McDermott KB. 2007. Generalizing test-enhanced learning from the laboratory to the classroom. *Psychon. Bull. Rev.* 14(2):200–6
McDaniel MA, Thomas RC, Agarwal PK, McDermott KB, Roediger HL III. 2013. Quizzing in middle-school science: successful transfer performance on classroom exams. *Appl. Cogn. Psychol.* 27(3):360–72
McDermott KB, Agarwal PK, D'Antonio L, Roediger HL III, McDaniel MA. 2014a. Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *J. Exp. Psychol. Appl.* 20(1):3–21
McDermott KB, Arnold KM, Nelson SM. 2014b. The testing effect. In *The SAGE Handbook of Applied Memory*, ed. TJ Perfect, DS Lindsay, pp. 183–200. London: SAGE
Melton AW. 1963. Implications of short-term memory for a general theory of memory. *J. Verbal Learn. Verbal Behav.* 2:1–21
Meyer AND, Logan JM. 2013. Taking the testing effect beyond the college freshman: benefits for lifelong learning. *Psychol. Aging* 28(1):142–47

Minear M, Coane JH, Boland SC, Cooney LH, Albat M. 2018. The benefits of retrieval practice depend on item difficulty and intelligence. *J. Exp. Psychol. Learn. Mem. Cogn.* 44(9):1474–86

Morris CD, Bransford JD, Franks JJ. 1977. Levels of processing versus transfer appropriate processing. *J. Verbal Learn. Verbal Behav.* 16(5):519–33

Morris PE, Fritz CO. 2000. The name game: using retrieval practice to improve the learning of names. *J. Exp. Psychol. Appl.* 6:124–29

Nelson SM, Arnold KM, Gilmore AW, McDermott KB. 2013. Neural signatures of test-potentiated learning in parietal cortex. *J. Neurosci.* 33(29):11754–62

Nunes LD, Karpicke JD. 2015. Retrieval-based learning: research at the interface between cognitive science and education. In *Emerging Trends in the Social and Behavioral Sciences*, ed. RA Scott, SM Kosslyn, pp. 1–16. Hoboken, NJ: Wiley

Pan SC, Pashler H, Potter ZE, Rickard TC. 2015. Testing enhances learning across a range of episodic memory abilities. *J. Mem. Lang.* 83:53–61

Pan SC, Rickard TC. 2018. Transfer of test-enhanced learning: meta-analytic review and synthesis. *Psychol. Bull.* 144(7):710–56

Pastötter BB, Bäuml K-HT. 2014. Retrieval practice enhances new learning: the forward effect of testing. *Front. Psychol.* 5:286

Putnam AL, Roediger HL III. 2013. Does response mode affect amount recalled or the magnitude of the testing effect? *Mem. Cogn.* 41(1):36–48

Pyc MA, Rawson KA. 2009. Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *J. Mem. Lang.* 60(4):437–47

Robey A. 2019. The benefits of testing: individual differences based on student factors. *J. Mem. Lang.* 108:104029

Roediger HL III, Agarwal PK, McDaniel MA, McDermott KB. 2011a. Test-enhanced learning in the classroom: long-term improvements from quizzing. *J. Exp. Psychol. Appl.* 17(4):382–95

Roediger HL III, Karpicke JD. 2006a. Test-enhanced learning: taking memory tests improves long-term retention. *Psychol. Sci.* 17(3):249–55

Roediger HL III, Karpicke JD. 2006b. The power of testing memory: basic research and implications for educational practice. *Perspect. Psychol. Sci.* 1(3):181–210

Roediger HL III, Karpicke JD. 2018. Reflections on the resurgence of interest in the testing effect. *Perspect. Psychol. Sci.* 13(2):236–41

Roediger HL III, Marsh EJ. 2005. The positive and negative consequences of multiple-choice testing. *J. Exp. Psychol. Learn. Mem. Cogn.* 31(5):1155–59

Roediger HL III, McDermott KB. 1995. Creating false memories: remembering words not presented in lists. *J. Exp. Psychol. Learn. Mem. Cogn.* 21(4):803–14

Roediger HL III, Putnam AL, Smith MA. 2011b. Ten benefits of testing and their applications to educational practice. In *Psychology of Learning and Motivation*, Vol. 55, ed. JP Mestre, BH Ross, pp. 1–36. Amsterdam: Elsevier

Rogalski Y, Altmann LJP, Rosenbek JC. 2014. Retrieval practice and testing improve memory in older adults. *Aphasiology* 28(4):381–400

Rohrer D, Taylor K, Sholar B. 2010. Tests enhance the transfer of learning. *J. Exp. Psychol. Learn. Mem. Cogn.* 36(1):233–39

Rowland CA. 2014. The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychol. Bull.* 140(6):1432–63

Rowland CA, DeLosh EL. 2015. Mnemonic benefits of retrieval practice at short retention intervals. *Memory* 23(3):403–19

Runquist WN. 1983. Some effects of remembering on forgetting. *Mem. Cogn.* 11(6):641–50

Schwieren J, Barenberg J, Dutke S. 2017. The testing effect in the psychology classroom: a meta-analytic perspective. *Psychol. Learn. Teach.* 16(2):179–96

Spitzer HF. 1939. Studies in retention. *J. Educ. Psychol.* 30(9):641–56

Szpunar KK, McDermott KB, Roediger HL III. 2008. Testing during study insulates against the buildup of proactive interference. *J. Exp. Psychol. Learn. Mem. Cogn.* 34(6):1392–99

Thomas RC, Weywadt CR, Anderson JL, Martinez-Papponi B, McDaniel MA. 2018. Testing encourages transfer between factual and application questions in an online learning environment. *J. Appl. Res. Mem. Cogn.* 7(2):252–60

Thompson CP, Wenger SK, Bartling CA. 1978. How recall facilitates subsequent recall: a reappraisal. *J. Exp. Psychol. Hum. Learn. Mem.* 4:210–21

Toppino TC, Cohen MS. 2009. The testing effect and the retention interval: questions and answers. *Exp. Psychol.* 56(4):252–57

Tse C-S, Balota DA, Roediger HL III. 2010. The benefits and costs of repeated testing on the learning of face-name pairs in healthy older adults. *Psychol. Aging* 25(4):833–45

Underwood BJ. 1972. Are we overloading memory? In *Coding Processes in Human Memory*, ed. AW Melton, E Martin, pp. 1–23. Washington, DC: Winston & Sons

Unsworth N. 2019. Individual differences in long-term memory. *Psychol. Bull.* 145(1):79–139

van den Broek GSE, Takashima A, Segers E, Fernández G, Verhoeven L. 2013. Neural correlates of testing effects in vocabulary learning. *NeuroImage* 78:94–102

van den Broek GSE, Takashima A, Wiklund-Hörnqvist C, Karlsson Wirebring L, Segers E, et al. 2016. Neurocognitive mechanisms of the "testing effect": a review. *Trends Neurosci. Educ.* 5(2):52–66

Vestergren P, Nyberg L. 2014. Testing alters brain activity during subsequent restudy: evidence for test-potentiated encoding. *Trends Neurosci. Educ.* 3(2):69–80

Wagner AD, Schacter DL, Rotte M, Koutstaal W, Maril A, et al. 1998. Building memories: remembering and forgetting of verbal experiences as predicted by brain activity. *Science* 281:1188–91

Watkins MJ. 1975. Inhibition in recall with extralist "cues." *J. Verbal Learn. Verbal Behav.* 14:294–303

Wing EA, Marsh EJ, Cabeza R. 2013. Neural correlates of retrieval-based memory enhancement: an fMRI study of the testing effect. *Neuropsychologia* 51(12):2360–70

Wixted JT. 2004. The psychology and neuroscience of forgetting. *Annu. Rev. Psychol.* 55:235–69

Yonelinas AP. 2002. The nature of recollection and familiarity: a review of 30 years of research. *J. Mem. Lang.* 46(3):441–517

Zerr CL, Berg JJ, Nelson SM, Fishell AK, Savalia NK, McDermott KB. 2018. Learning efficiency: identifying individual differences in learning rate and retention in healthy adults. *Psychol. Sci.* 29(9):1436–50

Annual Review of Psychology

Volume 72, 2021

Contents

Active Forgetting: Adaptation of Memory by Prefrontal Control
Michael C. Anderson and Justin C. Hulbert 1

"Reports of My Death Were Greatly Exaggerated": Behavior Genetics in the Postgenomic Era
K. Paige Harden 37

The Psychology of Reaching: Action Selection, Movement Implementation, and Sensorimotor Learning
Hyosub E. Kim, Guy Avraham, and Richard B. Ivry 61

Transcranial Magnetic Stimulation and the Understanding of Behavior
David Pitcher, Beth Parkin, and Vincent Walsh 97

Memory and Sleep: How Sleep Cognition Can Change the Waking Mind for the Better
Ken A. Paller, Jessica D. Creery, and Eitan Schechtman 123

The Cultural Foundation of Human Memory
Qi Wang 151

Trade-Offs in Choice
Franklin Shaddy, Ayelet Fishbach, and Itamar Simonson 181

The Origins and Psychology of Human Cooperation
Joseph Henrich and Michael Muthukrishna 207

Language as a Social Cue
Katherine D. Kinzler 241

Intergenerational Economic Mobility for Low-Income Parents and Their Children: A Dual Developmental Science Framework
Terri J. Sabol, Teresa Eckrich Sommer, P. Lindsay Chase-Lansdale, and Jeanne Brooks-Gunn 265

Moral Judgments
Bertram F. Malle 293

Integrating Models of Self-Regulation
Michael Inzlicht, Kaitlyn M. Werner, Julia L. Briskin, and Brent W. Roberts 319

The Psychology of Moral Conviction
Linda J. Skitka, Brittany E. Hanson, G. Scott Morgan, and Daniel C. Wisneski 347

Social Influence and Group Identity
Russell Spears 367

Socioeconomic Status and Intimate Relationships
Benjamin R. Karney 391

Experimental Games and Social Decision Making
Eric van Dijk and Carsten K.W. De Dreu 415

The Social Neuroscience of Prejudice
David M. Amodio and Mina Cikara 439

Psychology of Transnational Terrorism and Extreme Political Conflict
Scott Atran 471

Prejudice and Discrimination Toward Immigrants
Victoria M. Esses 503

Prejudice Reduction: Progress and Challenges
Elizabeth Levy Paluck, Roni Porat, Chelsey S. Clark, and Donald P. Green 533

The Science of Meaning in Life
Laura A. King and Joshua A. Hicks 561

Psychological Underpinnings of Brands
Richard P. Bagozzi, Simona Romani, Silvia Grappi, and Lia Zarantonello 585

Practicing Retrieval Facilitates Learning
Kathleen B. McDermott 609

Life Change, Social Identity, and Health
Catherine Haslam, S. Alexander Haslam, Jolanda Jetten, Tegan Cruwys, and Niklas K. Steffens 635

Stress and Health: A Review of Psychobiological Processes
Daryl B. O'Connor, Julian F. Thayer, and Kavita Vedhara 663

Understanding Human Cognitive Uniqueness
Kevin Laland and Amanda Seed 689

Psychology as a Historical Science
Michael Muthukrishna, Joseph Henrich, and Edward Slingerland 717

Indexes

Cumulative Index of Contributing Authors, Volumes 62–72 751

Cumulative Index of Article Titles, Volumes 62–72 756