HHS Public Access

Author manuscript

J Exp Psychol Anim Learn Cogn. Author manuscript; available in PMC 2023 July 01.

Published in final edited form as:

J Exp Psychol Anim Learn Cogn. 2022 July ; 48(3): 222–241. doi:10.1037/xan0000323.

Testing Improves Performance as Well as Assesses Learning: A Review of the Testing Effect with Implications for Models of Learning

Cody W. Polack,

Ralph R. Miller

State University of New York at Binghamton

Abstract

Taking a test of previously studied material has been shown to improve long-term subsequent test performance in a large variety of well controlled experiments with both human and nonhuman subjects. This phenomenon is called the *testing effect*. The promise that this benefit has for the field of education has biased research efforts to focus on applied instances of the testing effect relative to efforts to provide detailed accounts of the effect. Moreover, the phenomenon and its theoretical implications have gone largely unacknowledged in the basic associative learning literature, which historically and currently focuses primarily on the role of information processing at the time of acquisition while ignoring the role of processing at the time of testing. Learning is still widely considered to be something that happens during initial training, prior to testing, and tests are viewed as merely assessments of learning. However, the additional processing that occurs during testing has been shown to be relevant for future performance. The present review offers an introduction to the historical development, application, and modern issues regarding the role of testing as a learning opportunity (i.e., the testing effect). We conclude that the testing effect is seen to be sufficiently robust across tasks and parameters to serve as a hallmark phenomenon that theories of learning would best address. Our hope is that this review will inspire new research, particularly with nonhuman subjects, aimed at identifying the basic underlying mechanisms which are engaged during retrieval processes and will fuel new thinking about the learning-performance distinction.

Keywords

retrieval practice; learning/performance distinction; rehearsal; covert practice

The experimental study of learning dates back more than a century (see Thorndike, 1964/1898, for early research on instrumental learning, and Ebbinghaus, 1885/1964, for early research on associative learning which includes Pavlovian conditioning when the outcome has biological significance). In the early days of research on learning, the focus was on the experience needed for learning to occur, with the change in performance that

Mailing Address: Ralph R. Miller, Department of Psychology, State University of New York – Binghamton, Binghamton, NY 13902-6000, USA, TEL: (607) 777-2291, rmiller@binghamton.edu.

resulted from the experience being little more than a means of assessing what had been learned. That is, the focus was on the conditions that prevailed during training, with little or no interest in the information processing that must also occur at the time of testing in order for acquired memories to be expressed or the subsequent impact on behavior following the test. This started to change with Tolman's (1932) emphasis of the *learning/performance* distinction. Tolman's earliest demonstration of this distinction concerned the importance of proper motivation at the time of testing. Rats exposed to a maze exhibited no learning about the design of the maze devoid of food until they were later exposed to food in the goal box and then placed in the start box while food deprived. That Tolman's rats had learned the structure of the maze was evident in the rapidity with which they reached the goal box at test, compared to motivated rats that lacked prior experience with the maze. That Tolman's rats had obviously learned something to help them navigate the maze but not expressed it until they were properly motivated gave rise to the term *latent learning*, which refers to learning that in not observed during training, but is evidenced on a later test (*sensory preconditioning* is an example of an analogous phenomenon in the Pavlovian domain). In the present paper, we discuss the *testing effect*, which like heightened motivation, is an example of previously acquired performance being enhanced without additional training on the association in question. Adding test trials rather than additional acquisition trials can improve performance under some conditions. Thus, the testing effect is another example of the learning/performance distinction.

Various strategies have been proposed to improve retention and expression of learned materials. One of the most common is simple repetition, that is, more training (i.e., study) trials. After an initial study trial, any benefit from additional study trials could reflect new learning based on the additional presentations of the to-be-learned information and/or enhanced [subsequent] retrieval resulting from having retrieved the target information from reference memory. The benefit to [final] test performance following retrieval practice, relative to a control condition that receives further study trials in place of retrieval practice, constitutes the strongest demonstration of the *testing effect* (e.g., Karpicke & Roediger, 2007); however, any improvement in performance as a result of prior retrieval of the target information is commonly viewed as a manifestation of the testing effect. The testing effect is distinguished by its robust effect size and almost ubiquitous occurrence across many situations and a broad range of parameters (e.g., Roediger, Agarwal, Kang, & Marsh, 2009; Yang, Luo, Vadillo, Yu, & Shanks, 2021). In fact, among learning phenomena, it is perhaps second only to the trial spacing effect in being observed across diverse situations. Historically, testing was viewed as the key to assessing past learning. But the testing effect makes clear that testing is not only the means of assessing learning, it is in its own right an important tool for retention. The benefit of testing effect, that is, retrieval practice, is perhaps more surprising to the researcher who has been extensively exposed to the extinction literature than to a layman who lacks such knowledge concerning extinction but is highly familiar with using covert or overt rehearsal to commit information to memory.

Retrieval practice with feedback is ordinarily superior to retrieval practice without feedback. But even retrieval trials without feedback are seen to be more beneficial to subsequent performance than additional study trials (e.g., Butler & Roediger, 2008; Potts & Shanks, 2014). In terms of theory, this last observation is important because it precludes new

learning that would likely result from additional intact learning trials. If one defines *learning* as receiving and retaining information from the external world, then the testing effect is an instance of improved performance without new learning, relative to an identically treated control group that lacks only the retrieval-practice trials. In contrast, some researchers have viewed the improved performance seen as a result of retrieval practice (without feedback) as additional learning. They are using a different definition of *learning* than we are using; we have no argument with them, as one should not argue about definitions.

Although the testing effect is ordinarily described as enhance performance on a final test as a result of practice tests between target training and the final test, it sometimes takes the form of protecting the target memory from attenuation that is otherwise seen as a result of increasing retention intervals (particularly in proactive interference situations) and changes in context between target training and the final test (e.g., Pierce, Gallo, & McCain, 2017). A related effect is the test-potentiated new learning effect (aka the forward testing effect), which, as the name indicates, speaks to improved learning following tests [of related material] (Bjork & Storm, 2011; Chan, Meissner, & Davis, 2018). To the extent that the enhanced new learning differs from the target material, the phenomenon is arguably distinctly different from the conventional testing effect. But because it often contributes to the facilitated performance observed after prior tests, we will return to this issue in later sections.

Empirical investigations of the benefits of practice tests

The benefits provided by taking a practice test instead of further studying of the material have received a considerable attention in recent years. However, the basic premise, that practice recalling information improves later recall, is far from novel. The first experimental evidence suggesting the benefits of recall as a memory aid was provided by Abbott (1909). This early study lacked many of the methodological merits expected of contemporary research used today (e.g., the conclusions relied on introspection and a very limited sample), but it still stands as one of the earliest arguments for the benefits of retrieval practice.

Similarly, Gates (1917) publicized the benefits of testing by demonstrating that self-testing was an effective performance enhancing strategy. Students from diverse grade levels were asked to covertly recite information (either nonsense syllables or short biographies) after being allowed to read them for a short time. In addition to looking across multiple developmental age ranges (Grades 1, 3, 4, 5, 6, & 8), Gates manipulated the percentage of time during the fixed duration training sessions that the children spent self-testing (i.e., covert recitation) items (nonsense syllables or biographical information) using six different fractions of the session duration (varying from 0–90%) for retrieval practice with the remainder of the session duration having been allocated for initial study, as well as performing both immediate and delayed (up to 3 hours) assessments, thus generating a 6 (age) × 2 (information type) × 6 (retrieval time %) × 2 (retention interval) design. In short, this impressive undertaking demonstrated that every age group except the youngest (1st graders) demonstrated some benefit from additional time spent performing retrieval practice. Moreover, the benefit of covert recitation was observed whether the learned information was in the form of nonsense syllables or biographical information. However, performance on the

biographical information reached asymptote and actually began to drop off after the children practiced retrieval for over 60% of the session duration (see Roediger & Karpicke, 2006b, Figure 1). This observation suggests that in this specific situation sacrificing study time for practice testing becomes counterproductive below 40% study time. Gates' study was the first to reflect the necessity of sufficient acquisition for retrieval practice to enhance retention. However, it should be noted that McDermott and Naaz (2014) failed to replicate the basic findings of Gates' influential study when they used adults rather than children.

Approximately twenty years later, a second industrious undertaking produced another large-scale example of the testing effect in school aged children. Spitzer (1939) had 3605 6th graders read two different passages, followed immediately by a four-item multiple choice test covering one of the two passages. Two groups of children then received an immediate multiple-choice test covering the second (i.e., non-tested) passage, whereas six other groups of children received this test at later dates, ranging from 1 to 63 days. Two weeks after these practice tests, all the delay groups received a second identical test. The two immediate groups received additional testing either 1 or 7 days later and then a final test 2 weeks or 8 weeks later to provide some controls for initial learning and retention over the length of the entire study. Comparisons among the delay conditions allowed Spitzer to conclude that delaying retrieval practice produced lower test scores on the first as well as the second test. It is interesting to note that these students never received feedback on their performance for these practice tests. Thus, one might expect errors on the first test to perseverate to the second test. Using a limited sample of the collected data, Spitzer discovered that students perseverated correct answers 79% of the time and errors in half of the cases. If one takes into consideration that guessing would net 25% consistency between errors, this finding suggests there is the potential for testing to preserve false recollections (i.e., negative testing effect; Roediger & Marsh, 2009) in addition to successful recall. The positive aspect of this finding is that the repetition of errors was less common than the repetition of correct answers, indicating that the retrieval benefit to erroneous responses may be somewhat more limited than the corresponding benefits to correct answers offered by testing. We will return to this topic later in our discussion of the role of feedback during retrieval practice.

These early studies of the testing effect provided the foundation for further investigations; however, the modern surge in interest didn't gather momentum until the late 80's and early 90's when serious attention to the benefits of testing were raised. One explanation of the testing effect is that an additional study trial seemingly requires less mental effort than a practice test trial. The assumption was that simply reading the information again requires a relatively superficial amount of processing relative to taking a practice test in which the material is considered and an answer must be generated (e.g., Craik & Tulving, 1975; Kolers, 1973). However, Glover (1989) argued that the basic testing effect is not determined simply by the relatively large amount of processing that occurs during a test. By varying the interval between initial study and the practice test, Glover (Experiment 2) suggested that amount of processing would be preserved across the two identical tests; therefore, such an account would expect equivalent benefits (see also, Kane & Anderson, 1978). When practice testing occurred immediately after the study period, performance on the final test was diminished relative to when the practice test occurred two days after the study phase,

suggesting that the amount of processing per se that occurs during the practice test is insufficient for explaining the influence of such delays between study and practice testing.

At first inspection, Glover's (1989) observation of a benefit of a delayed over an immediate practice test appears to be in stark contrast with the previous observations of Spitzer (1939). Spitzer observed a decrease in final performance as retrieval practice was delayed, whereas Glover observed a benefit of delayed retrieval practice. However, a number of differences between the two preparations begins to explain this discrepancy. Spitzer's findings appear to be the result of a reduction in retrieval success between initial study and the practice test due to longer delays with younger participants, whereas Glover (1989) reports no such differences in retrieval failure on the first test resulting from delayed retrieval practice with his parameters, that is, shorter delays and college aged participants. Additionally, Spitzer's practice test did not include feedback. The inclusion of feedback on the practice test in Glover's study may have compensated for greater retrieval failure when the first test was delayed. Any number of these factors could have contributed to differences in successful retrieval during practice testing, which could have resulted in differential amounts of retrieval practice actually occurring. Each of these factors could play a potential role in determining the benefit provided by a practice test and they represent some of the key variables of interest in the current literature.

An additional point of note in addressing differences in the finding of Spitzer (1939) and Glover (1989) is that Glover's design confounded the delay between initial study and the practice test with the delay between the practice test and final assessment. Thus, the benefit to final test performance following delayed practice testing may have been due to the shorter interval between the practice test and final assessment. In light of this possibility, it might seem that differential amounts of processing should be retained as a possible source of the testing effect. However, McDaniel and Fisher (1991) provided further evidence against the amount-of-processing account by demonstrating that deeper processing of feedback during practice testing had little influence on the testing effect. When learners were asked to elaborate on the feedback they received during a practice test, their recall performance was no better than if they merely read the feedback. But, given that this finding was predicated on a null result, it too should be accepted with caution.

Glover (1989) initiated much of the current interest in the testing effect by providing a particular cognitive account of the effect. Specifically, he argued that the benefit of retrieval practice was a function of how complete the retrieval event was and the number of successful retrievals during the retrieval practice period, thereby, paving the way for new cognitive theories regarding human memory (e.g., Bjork & Bjork, 1992). The central suggestion is that, once target information has been retrieved, the information is subsequently more readily retrieved owing to either a strengthening of the memory or, as Bjork and Bjork (1992) suggest, enhancement of a retrieval process for existing associations that is separate from the target association itself.

The studies described above (Gates, 1917; Glover, 1989; Spitzer, 1939) suffer from a number of procedural flaws. Perhaps, most important is that none of the described experiments controlled for mere exposure to the material by comparing retrieval practice

groups to groups that received additional study time (Skaggs, 1920). Carrier and Pashler (1992) provided an early example that a testing benefit can be observed using conservative controls for exposure. Such controls are critical because practice tests constitute additional exposure to at least parts of the studied information when feedback is absent and all the information when feedback is present. This additional exposure provides an opportunity for additional study. Therefore, a more conservative experimental design would include a control condition with the opportunity for additional study of the material in place of the practice tests. Carrier and Pashler included such a control condition in a within-subject design and demonstrated that providing additional studying time to the control condition does not eliminate the testing effect although it does diminish it.

In a paired associate task containing Eskimo-English word pairs, Carrier and Pashler (1992; Experiment 3) presented undergraduate participants with a list of 40 word-pairs in serial fashion. Participants were exposed to each word pair twice. Following this initial training phase, half the word pairs were presented to the subject for a third time (Study condition), whereas the other half of the items were presented as practice test trials (Test condition). The practice test trials consisted of presentations of only the Eskimo word and the participant being given a limited time to recall the English associate before the answer was provided (i.e., feedback). Even using this conservative control, the Test condition yielded better performance than the Study condition at the final assessment. This difference was observed following either a 5-min or a 24-hour retention interval. Carrier and Pashler's paper marks the beginning of modern methodology in assessing the benefits of testing and provide one of the first well controlled demonstrations of testing providing a benefit over mere exposure (see also, McDaniel & Fisher, 1991).

Application to real-world settings

Much of the interest in the benefits of testing as a mnemonic aid has been due to the implications for educational reform (for an extended discussion, see Roediger et al., 2010). Although 'retrieval practice' directly increasing retrieval strength is the commonplace explanation of the testing effect, a number of additional benefits of practice tests (that might actually underly the presumed enhancement of retrieval strength) have been suggested that emphasize changes in organization, familiarity with test format, and study strategies (Roediger, Putnam & Smith, 2011). Unfortunately, students overwhelmingly prefer to review lecture notes rather than take a practice test and they feel more confident in their performance after doing so (Karpicke, Butler, & Roediger, 2009; but see Tullis, Fiechter, & Benjamin, 2017). Although many of the findings regarding the testing effect have been obtained in laboratory settings using stimuli that arguably lack ecological validity (e.g., memorizing arbitrary word lists; Hogan & Kintsch, 1971; Izawa, 1967, 1970; McDaniel & Masson, 1985; Thompson, Wenger & Bartling, 1978; Tulving, 1967; Wheeler, Ewes, & Buonanno, 2003), there has been growing interest in the use of testing to improve retention in settings closer to real world situations such as understanding and applying conceptual information, rather than rote memorization (Butler, 2010). For example, McDaniel and Fisher (1991) extended the testing effect to an incidental learning paradigm with random trivia. Butler, Karpicke, and Roediger (2007) similarly demonstrated the testing effect using information pertaining to basic knowledge questions.

Word lists can be useful in some limited settings and paired associates may offer some benefit to early second language learning when students must first map novel foreign words into their primary language. However, to be truly useful in educational settings the information learned must generalize from training to formal testing (i.e., the proximate goal of education) as well as from training to life outside the classroom (i.e., the ultimate goal of education). A few examples of studies with a high degree of ecological validity come from the classic experiments already described in which participants read passages during training and then used multiple choice (Spitzer, 1939) or free recall (Glover, 1989) for practice tests. Similarly, Gates (1917) used short biographical passages with covert recitation in place of practice testing, an unorthodox retrieval practice by present standards. It should be noted that covert retrieval practice appears to be no less effective than the overt practice in terms of the size of subsequent retrieval benefits (Smith, Roediger, & Karpicke, 2013; but see, Jönsson et al., 2013).

Further studies of ecological validity have gone so far as to assess the testing effect in an experimental college classroom setting. Butler and Roediger (2007) compared studying lecture notes immediately following a lecture to taking either a multiple-choice test or short essay test. Somewhat surprisingly, only the short essay test produced any observable increase in performance on the final test relative to the study control. What is notable here is that, although the presentation format of information during the initial test phase was qualitatively different from that of the initial training phase (i.e., lecture), the benefits of practice testing remained. The testing effect has also been evaluated during both traditional and online college courses, demonstrating that this can be a valuable and flexible tool for improving student performance in both conditions (Logan, Thompson, & Marshak, 2011; McDaniel, Wildman, & Anderson, 2012).

In a similar vein, Johnson and Mayer (2009) provided evidence for a testing benefit when the training phase was an audio-visual presentation which described the formation of lightning. The study-only control subjects simply watched the presentation again, whereas two different practice test conditions were implemented. One practice test condition involved a written description of how lightning forms (retention test), and the second practice test condition asked for a more analytical response such as, “What could you do to decrease the intensity of lightning?” (i.e., a transfer test; p. 622). Their findings indicated that both practice test conditions provided a benefit at final recall. Not only did practice testing provide a benefit when there was a discrepancy between the training and test phase, but also the use of an analytical practice test produced a testing effect. It is important to note that a testing benefit was only observed when the final assessment matched the respective practice test. If subjects received the transfer test (i.e., a test that was in a different format from the practice test), their performance was no different from the study-only control on the final assessment. This suggests a potential role of transfer-appropriate processing in this preparation (Craik & Lockhart, 1972; Morris, Bransford, & Franks, 1977). Johnson and Mayer (2009) argued that transfer-appropriate processing may offer a full account of the testing effect, a possibility we discuss later on. In general, these findings support the benefit of retrieval practice (i.e., practice tests) for enhanced recall in a wide variety of real-world settings.

The testing effect has also been evaluated across a wide age range from early childhood to middle adulthood. Again, the early work of Gates (1917) provides an extensive cross-sectional data set for the influence of retrieval practice, notably demonstrating the benefit of covert recitation for all ages except their youngest participants (i.e., first graders). Several researchers have continued in the tradition of Spitzer (1939) to determine the applied value of testing in high school and elementary school populations in addition to the more common college settings described above (Carpenter, Pashler, & Cepeda, 2009; Glover, Krug, Hannon, & Shine, 2010; Marsh, Agarwal, & Roediger, 2009; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011). The testing effect has been demonstrated in children as young as second graders (Goswick, Fazio, & Marsh, 2010). Most researchers have focused on using school-aged participants. An interesting counterexample is provided by Tse, Balota, and Roediger (2010), who found that in older adults (70-year olds) repeated testing did not produce an improvement in recollection over a study control, whereas the elderly (80-year olds) actually showed a decrement following repeated testing relative to additional study. These results were only observed in the absence of feedback which potentially allowed practice testing to enhance retrieval of incorrect answers (Experiment 1). However, if corrective feedback was provided during the practice test, then the typical benefit of testing was observed in both age categories (Experiment 2). Their optimistically labeled “middle-aged” category (60-year olds) demonstrated the testing benefit both with and without feedback, indicating that there was not something idiosyncratic about their no-feedback condition that generally failed to improve retention. Given these findings, we can tentatively establish an effective range (between ages 7 and 70) in which testing without feedback provides a benefit. The interpretation offered by Tse et al. regarding their older sample is that the subjects in their 70s and 80s could not produce enough correct answers on the first practice test in order for the test to provide effective retrieval practice. In such circumstances, when memory for the learned material is limited, additional exposure to the full original material may be needed, whether it be in the form of additional study time or in the form of testing with feedback. This finding raises the question of whether first graders also failed to demonstrate a benefit from retrieval practice resulting from an inability to successfully retrieve the information during practice tests or from their being refractory with respect to the testing effect.

In applied setting, retrieval tests are usually accompanied by feedback, which makes it difficult to parse the degree that the resultant improvement in performance arises from improved retrieval as opposed to new learning. But the many demonstrations of the testing effect without feedback, strongly suggests that the testing effect contributes to the improved performance observed when there is feedback.

Modern test effect phenomena and boundary conditions

Generalization across types of tests.

More recent testing effect studies have attempted to identify its boundary conditions. One potential constraint on the benefits of testing is that benefits may only be observed when final testing uses the same procedure used during practice testing. This expectation is based on an account of the testing effect that hinges on context shifts, in which the task

demands of the final test serve as part of the test context. This orientation is emphasized by proponents of a transfer-appropriate processing account of the testing effect, which suggests that practice tests facilitate later performance because the processing that occurs during the practice tests is more similar to the processing that is to occur on the final test than is further studying. Therefore, changing the format of the test between retrieval practice and final assessment would be expected to reduce the benefits of testing (e.g., Johnson & Mayer, 2009).

Butler (2010) attempted to address whether the testing effect would persist in enhancing performance across different types of tests of the same material. Undergraduate participants were asked to study six different passages. These passages contained factual information as well as conceptual information based on Bloom's (1956) taxonomy. For each fact and concept, three different questions were generated to produce different wordings for each question, but the correct answers across the three variants were identical. Participants were first asked to read each passage. Following this study phase, participants were asked to reread two of the passages and take tests on the other four passages. Each passage was either reread or practice tested three times. Half of the practice tested passages used identical questions during each test, whereas the other two passages received novel wordings of the questions during each practice test. One week later at the final assessment, Butler tested participants on the material using the questions the participant had already seen during the practice test (Experiment 1a) or performed the final assessment using novel inferential questions (Experiment 1b). The restudy condition received comparable questions because the passages were appropriately counterbalanced across participants. The only difference between questions pertaining to the restudied passages and the practice tested passages was that the restudied passages had never before been tested in any format. Experiment 1a suggested that retention of the material was improved so long as testing of any sort occurred. Variability in how the questions were worded did not produce a reliable difference in responding from testing with the same wording each time. Additionally, the type of question (factual or conceptual) did not influence final performance. This pattern of findings was also observed in Experiment 1b, which assessed transfer of learning to a novel inferential test. Thus, retrieval practice appears to improve transfer, but increasing initial variability in testing does not seem to influence subsequent transfer. However, it is possible that the variability introduced during retrieval practice (i.e., wording differences) did not provide a benefit at later transfer to a conceptually different assessment (i.e., inferential questions) because the variability offered by changing words was a relatively superficial change in the type of question. There are many demonstrations in which training in multiple contexts assists in later retrieving that content in a novel setting (Dunsmoor, Fredrik, Zielinski, & LaBar, 2014; Glautier & Elgueta, 2009; Gunther, Denniston, & Miller, 1998; Laborda & Miller, 2013; Miguez, Laborda, & Miller, 2014; Vansteenwegen et al., 2007). Studies focused on increased transfer across physical contexts when training occurs in multiple contexts primarily have examined instances in which there was some source of potential interference between different associations rather than mere acquisition of a single association. This would suggest that the experimental design used by Butler could have could have been insensitive to context manipulations. Additional research would be necessary to specifically address this possibility.

In an unusual example of the testing effect, Goode, Geraci, and Roediger (2008) examined whether variability during training would facilitate later retrieval in a task involving anagrams. Their procedure entailed repeated exposure to different anagrams of the same word followed by testing on a novel anagram of the same word (Varied Condition) or repeated presentations of the same anagram followed by testing on the exact same anagram (Same Condition). They found that exposure to different arrangements of the same letters increased accuracy at identifying the anagram relative to repeating the same arrangement of letters each time. Goode et al. claimed that this demonstrated the influence of repeated testing on higher-order types of questions (i.e., anagrams), which makes it somewhat different from the typical testing effect. Additionally, the nature of their task omitted any explicit training vs. practice phase, which raises the possibility that the anagram practice improved the participants' ability to solve anagrams in general. Alternatively, their findings might be explained in terms of increased exposure to the correct letter placement across presentations of the different letter arrangements, rather than learning some general skill in solving anagrams. In the Varied condition each anagram contained some of the letters located in the correct place. Therefore, those participants who were exposed to multiple configurations likely were exposed to more of the letters in the correct place, which could have facilitated retrieval of the correct item.

Although a considerable amount of research has been aimed at illuminating how well the testing effect generalizes from one set of materials to another, Butler (2010) is one of the few examples to demonstrate that the testing effect generalizes from practice on one type of test to conceptually different tests (also see Carpenter, Pashler, & Vul, 2006; Johnson & Mayer, 2009; Rohrer, Taylor & Shalor, 2010). In a study by Halamish and Bjork (2011), practice testing in the form of a cued recall test resulted in a benefit on a subsequent free recall test; however, an advantage for additional study over practice testing with cued recall was observed when the final test was in the form of a cued recall test (Experiment 2). Increasing the match between the practice test and final assessment (both cued recall tests) in Halamish and Bjork (2011) actually hindered performance, which they concluded had to do with the difficulty of the final assessment rather than the encoding-retrieval match. Using their distribution framework account, they argue that an easier final assessment, like a cued-recall test, would benefit more from additional study than from cued-recall practice. But they still observed a retrieval-practice benefit when participants had to generalize to a more difficult final assessment.

Practice testing may also serve to bias the information attended to on future study trials and additional practice tests that precede the final test. This bias would likely depend on the specific format of the practice test. Bjork and Strorm (2011) gave a practice test in which students completed fill-in-the-blank questions based on previously learned information. Participants exposed to this type of practice test or a study-only condition were then given new learning materials. It was anticipated that those participants who were practice tested with a fill-in-the-blank test would be biased towards picking up on contextual information during the next training session. Critically, when tested on recall for contextual information alone with respect to the new learning materials, the participants who had been previously exposed to the fill-in-the-blank test demonstrated better retention for context information than the study-only participants. This point is critical, as it suggests that testing shapes how

future information will be processed. The task demands of the initial practice test improved performance on subsequent tests, presumably by altering the content of what participants attended to.

Timing in the testing effect

Temporal relationships play several critical roles in determining the testing effect. The interval between retrieval practice sessions has been extensively evaluated (e.g., Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Karpicke & Roediger, 2007). Throughout this research, it becomes clear that the interval between training and the first retrieval practice is a critical variable in producing a reliable benefit from testing. If the interval between initial training and retrieval is too long, then the risk that retrieval may be incomplete during retrieval practice increases (Fazio, Argawal, Marsh, & Roediger, 2010), with the result being a smaller benefit on the final test. If the interval is too short, then the information from training may still be active in short-term memory, making retrieval inconsequential. Presumably, one cannot practice retrieving information that is currently active because there is no information or limited additional information to retrieve back into an active state. This could lead to a situation in which the optimal benefit produced by practice testing is when these tests require the desirable amount of difficulty for retrieval. This ‘desirable difficulty’ is thought to occur when information is sufficiently difficult to recall yet complete retrieval is still possible (Bjork, 1994). The amount of delay prior to retrieval practice that is optimal is, in part, determined by the amount of delay between retrieval practice and final recall (Cepeda et al.). Specifically, when the interval between initial training and final recall was long, the optimal interval between training and retrieval practice was also longer. However, Cepeda et al. found that the optimal delay for retrieval practice was a smaller proportion of the total interval between training and final testing, the longer total retention interval was.

In addition to the interval between initial training and retrieval practice, the delay between the retrieval practice and testing has been shown to produce differential findings from comparable delays using only restudy (Roediger & Karpicke, 2006a). In their widely cited study, Roediger and Karpicke asked undergraduates to read prose passages on scientific topics. Practice test and restudy conditions were manipulated within-subjects by having participants study two different passages. Following initial training (i.e., reading the passages), participants were asked to restudy one passage and perform a written free recall on the other passage. The order of restudy and test was counterbalanced across subjects. Critically, the delay between the re-study/practice test phase and final assessment was manipulated using retention intervals of 5 min, 2 days, or 7 days. Their results are compelling because they clearly demonstrated an interaction between delay and whether the passage was restudied or practice tested. In fact, with the 5-min delay, the restudy condition exhibited a reliable benefit on final recall relative to the practice test condition, whereas after 2 or 7 days the practice test condition exhibited superior performance on the final test. In their Experiment 2, Roediger and Karpicke allowed multiple restudy or practice test opportunities, and then compared these to a hybrid of restudy and practice testing during the 5-min and 7-day retention intervals. The findings were again clear, with the 5-min delay, more study opportunities were monotonically related to better performance, whereas, after 7 days more practice test opportunities improved performance. It was suggested that

this finding supports consolidation as having a role in the testing effect. That is, benefits from retrieval practice may require that initial information first be stored before benefits of retrieval practice can be realized (Racsmany, Conway, & Demeter, 2010). Alternatively, this may be the result of differential rates of decay following the different types of initial learning, subsequent study, and testing (Wheeler et al., 2003). Congleton and Rajaram (2012) argue that the ability to organize the information during a free recall practice test leads to more relational processing of the information which would be more resistant to decay compared to item specific processing promoted by study conditions. Although there is clearly some influence of the interval between retrieval practice and final testing on the benefits of retrieval practice tests over additional study, some immediate benefits emerge in terms of latency to respond even on the immediate test itself (Kersztes et al., 2013; MacLeod & Nelson, 1984; van den Broek, Segers, Takashima & Verhoeven, 2013). The tendency for the testing effect to be weaker or nonexistent during an immediate test has also been shown to be dependent on the level of performance during retrieval practice and may be mitigated by providing adequate feedback during retrieval practice (Rowland & Delosh, 2014).

There has also been a significant amount of interest in the role of the intervals between successive retrieval practice tests. Much of the discussion on this topic has driven by the issue of whether retrieval practices should be evenly spaced or distributed along an expanding schedule (Landauer & Bjork, 1978). The theoretical motivation behind the expanding schedule is to provide greater opportunity for achieving the best level of desirable difficulty between retrievals, as retrieval becomes more difficult with longer delays (e.g., Ebbinghaus, 1885/1964). The assumption is that shorter intervals between retrieval practices prevent early retrieval failure when the participant is less facile with the material. Increasing the delays to later retrieval attempts to scale the delays of retrieval practice with the participant's increased ability to retrieve the target information (Cepeda et al., 2008; Landauer & Bjork, 1978).

In general, the introduction of a long interval between a participant's last experience with an item and subsequent retrieval makes that information harder to recall. The consequence is that a greater number of errors may occur during practice tests that are greatly delayed. If errors are made, then retrieval practice is not successful and should not provide a benefit to final testing. Indeed, errors during retrieval practice might be expected to actually impair later performance. Pashler, Zarow, and Triplett (2003) have looked at the spacing between practice tests with respect to the possibility that this interval between tests may influence the chance for errors to be made. They observed that longer delays between successive retrieval sessions increased errors committed during the retrieval practice session, but performance at the final test, one day or one week later, was enhanced relative to when a shorter delay occurred between practice tests. This finding suggests that although spacing between retrieval practice sessions increases initial errors, there remains an overall benefit of spacing retrieval sessions. However, it is likely that the use of corrective feedback in Pashler et al.'s design greatly limited the negative influence of errors present during the delayed retrieval practice, thereby allowing the effect of spacing to provide the dominant influence on responding.

Feedback and pretesting

Considerable attention in the testing effect literature has been directed at the role of feedback during the practice tests (e.g., Butler et al., 2008; Butler & Roediger, 2008; McDaniel & Fisher, 1991; Vojdanoska, Cranney, & Newell, 2010). It is widely accepted that feedback enhances the benefits of testing by allowing the participant to correct errors that were made during retrieval, thereby reducing the likelihood of the error from being repeated during a later test (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991). Additionally, feedback has been shown to improve performance even when the participant provides correct answers during retrieval practice by reinforcing the answer provided and increasing the learner's confidence in their response (Butler et al.).

Butler et al. (2008) administered a multiple-choice test covering general knowledge questions in the absence of any initial exposure or study phase. Each participant received feedback on half of the presented questions and no feedback on the other half. Participants were forced to answer each question, even if they had to guess, and then enter their confidence rating using a four-point scale. On occasions when feedback was provided, it occurred following the confidence rating. At the final test, all the items from the previous phase were presented along with a set of novel multiple-choice questions. Not surprisingly, when feedback was provided, the testing effect was more robust than when feedback was absent; however, even in the absence of feedback there was an observable benefit of practice testing compared to having no prior experience with the test items. What made Butler et al.'s findings so compelling was their assessment of how feedback influenced both correct and incorrect responses. The data were broken down in terms of the proportion of responses correct on the final test that had been either incorrect or correct during the initial test. When feedback was presented during practice testing, a greater proportion of the initial errors were corrected on the final test. Additionally, a greater proportion of correct answers on the initial test were repeated during the final test when feedback had been provided. Moreover, when reports of the participants' confidence were considered, it was determined that feedback selectively improved retention of correct answers that were rated with lower confidence ratings (see also, Smith & Kimble, 2010).

The greater benefit of practice testing with feedback makes intuitive sense considering that it does not interfere with retrieval practice; yet, it provides an opportunity for additional study during instances of retrieval failure. The benefit of testing in the absence of feedback is less certain, considering there is the opportunity to make an error and the association to the erroneous response could be strengthened. Those advocating for errorless learning in educational settings argue that the more errors that are emitted during initial tests, the more likely the participant is to commit that error in the future (Skinner, 1958). The benefits of the retrieval practice effect do not discriminate between correct or incorrect answers. That is, if one continually rehearses incorrect information, then that erroneous information will be more readily available during later tests. That perseveration of errors can occasionally result in an impairment in final recall resulting from practiced retrieval relative to additional study is known as the 'negative testing effect' (a.k.a. the reverse testing effect; Chan, Thomas, & Bulevich, 2009; Roediger & Marsh, 2005; Peterson & Mulligan, 2013), and is especially prevalent in multiple-choice tests which provide attractive lures. A series of experiments

providing students (college and high school) with practice SAT tests reported that if initial test performance was sufficiently low, final performance was actually hindered by practice tests (Marsh et al., 2009). Thus, the benefit of testing without feedback seemingly requires high confidence in addition to correct responses in order to be beneficial, otherwise incorrect responses are apt to perseverate.

The Butler et al. (2008) preparation described earlier deviates from the more typical study - practice test - retest paradigm for the testing effect by omitting the initial study phase. One might consider that, at least for general knowledge questions, participants' personal experience prior to the experiment constitutes a pre-experimental initial study phase. There has also been interest in the benefits of practice testing alone (without initial study) on later test performance when participants have little knowledge of the material prior to the experiment (i.e., the pretesting effect; Richland, Kornell, & Kao, 2009). In fact, to ensure that the study did not include material with which participants had prior knowledge, correct answers during pretesting (~5% on average) were removed from the final test on an individual basis. Therefore, feedback provided during practice testing provided the first experimental exposure to the correct material. The implication of the observed benefit provided by pretesting cannot easily be explained in terms of retrieval practice per se because the participants did not yet have experience with the target information during the pretest because the pretest answers participants provided were all incorrect. Thus, it is argued that something about the testing format in these experiments enhances processing of the answer when it is first encountered. Pan and Sana (2021) demonstrated that pretesting benefits can be larger than those of posttesting and these benefits appear to be gained from increased processing during subsequent experience with the test material. This finding is surprising considering the quantity of errors that are bound to be made on an initial test, unless the practice tests are made so easy as to be uninformative. According to an errorless learning approach, the increased chance of making an error should limit any testing benefit provided by pretesting and increase the likelihood of a negative testing effect. The presence of feedback alleviates this concern by presenting the correct response when the learner is unable to retrieve the appropriate response on their own, but it is crucial to recognize the influence pretesting has on future information processing during study trials.

There is some limited support for the alternative claim that learners actually benefit from committing errors during a pretest containing feedback (Cunningham & Anderson, 1968; Izawa, 1967, 1970; Kane & Anderson, 1978; Kornell, Hays, & Bjork, 2009). The observation that errors improve learning has not been shown in the traditional design of study followed by practice tests but rather in preparations like that used by Butler et al. (2008) in which participants are initially tested without a study phase. Due to the interest of Kornell et al. in errors produced during initial testing, their preparation had to ensure that participants would get numerous items incorrect on initial testing. To achieve this end, they presented subjects with fictitious questions (e.g., asking for the author of a nonexistent book) interspersed with actual general knowledge questions. The general knowledge questions were included solely to make the fictional questions less conspicuous (Berger, Hall, & Bahrick, 1999). In the basic preparation, participants were exposed to half the questions, orthogonal to whether the questions were fictitious or factual, in the form of initial test trials, whereas the other half of the questions were presented as study-only trials. Test trials

consisted of presenting the question alone for 8 s on a computer screen, during which time the participant was asked to use a keyboard to type the answer. Each question-answer pair was presented only once. Following the 8-s presentation, the answer was provided in addition to the question for an additional 5 s. Study-only trials simply presented the question in compound with the answer for either 5 s (Experiment 1) or 13 s (Experiment 2). Following a short delay, subjects were tested on all of the items again, this time in the absence of any feedback. The results indicated that when total exposure to the answer was matched (Experiment 1), participants performed better on the fictitious questions on which they had been tested previously, even though they were sure to make errors during the initial assessment. However, when the duration of the study-only trials was matched to the entire duration of the test trial (Experiment 2), no difference was observed between the two conditions. Kornell et al. concluded that their findings still support a benefit of test trials because performance was no different from the study-only trials, which provided over twice as much exposure to the answer. In a later experiment (Experiment 6), items for which participants actually entered an incorrect response were found more likely to be correct on the final test than when the participants had omitted the response. This last observation supports the counterintuitive conclusion that emitting an error is actually more beneficial for later accurate recall than not selecting an answer, at least when corrective feedback is provided. Seemingly, correction led to inhibition of the erroneous answer. Experiment 6 differed from Experiments 1 and 2 in that weak associates (e.g., Tree-Sycamore) were used instead of fictitious questions, and treatments were manipulated between-subjects rather than within. Knight, Ball, Brewer, DeWitt, and Marsh (2012) replicated the findings of Kornell et al., but found that this benefit was seen only when there was some semantic relationship between the cue and target information. When unrelated paired associates were used, incorrect initial retrieval significantly impaired performance relative to a study-only control. This observation led Knight et al. to suggest that guessing incorrect but related material serves to mediate learning the target information. Potts and Shanks (2014) found that this semantic relationship is not crucial and pre-testing can provide a benefit even when the paired associates are entirely novel. Further analyzing only explicitly incorrect generation at pretest, Potts and Shanks (2014) were still able to observe a mnemonic benefit.

Given the evidence reviewed thus far, it may seem that feedback is universally beneficial during testing. However, there are certain situations in which feedback provides no benefit or even attenuates the benefit of practice tests. For example, the benefit of feedback has been to be dependent on task difficulty, such that when the retrieval of the correct response is easy, feedback tends to be irrelevant (Hattie & Timperley, 2007; Pyc & Rawson, 2009). Clearly, the benefits offered by feedback depend on both the nature of the task as well as the quality of the feedback. Benefits of feedback are observed most consistently in rote memorization tasks when the feedback comes in the form of providing the correct answer. When feedback is limited to simply indicating whether the answer was correct or incorrect, no benefit of feedback is observed (Fazio, Huelser, Johnson, & Marsh, 2010; Pashler, Cepeda, Wixted, & Rohrer, 2005). Kantner and Lindsay (2010) argue that the benefit of feedback is especially limited in recognition tasks. A benefit of feedback also appears to be more difficult to observe in skill-based tasks (Rosenbaum, Carlson, & Gilmore, 2001; Schmidt, Young, Swinnen & Shapiro, 1989). Additionally, in circumstances when time is

limited prior to a final test, feedback costs the learner valuable time that might be better spent on performing retrieval practice with additional test items.

Hays, Kornell, and Bjork (2010) found that when reviewing feedback takes away from time that could be spent performing additional retrieval practice, feedback actually becomes detrimental to final performance. Hays et al. presented undergraduates with two different lists of foreign word-English word paired associates. After initial exposure to the list of associates, participants were presented with the foreign word alone and asked to recall its English associate. During this retrieval practice, each participant received feedback after each response for one list, but for the other list feedback was occasionally omitted. The time spent during this retrieval practice phase was strictly controlled, such that when feedback was omitted the participant was able to perform an additional retrieval trial from that list of items. There were two ways in which items could have been skipped. Some participants were allowed to decide when to skip feedback, whereas for others feedback was automatically skipped whenever a correct response was made. Hayes et al. anticipated that omitting feedback to receive additional retrieval practice would produce better retention of the material. Additionally, due to the finding that undergraduates undervalue the benefit of retrieval practice (Karpicke et al., 2009; Kornell & Son, 2009), participant-controlled omission of feedback was anticipated to be suboptimal compared to when omission was automatic. Omission of feedback was found to provide a benefit overall, indicating that, at least when feedback prevents additional retrieval practice, there is a cost to providing feedback that indirectly affects final test performance by removing opportunities for further retrieval practice. Trivially, the difference between automatic and participant-controlled omission of feedback was not reliable, meaning that, although participants did not skip feedback *optimally* as defined by the authors, their performance was comparable to the optimal skipping condition.

Testing effect in nonhuman animals

Humans are not the only species to benefit from retrieval practice of previously learned information. In the 1980's a number of reports indicated that presentations of a conditioned stimulus (CS) presented alone could serve as a reminder of a previously learned association with an unconditioned stimulus (US). These reminder treatments included brief presentations of the CS-alone, which is identical to how one would test an animal for Pavlovian conditioning. Many of the studies used reminder treatments to recover a memory from the effects of retrograde amnesia (e.g., DeVietti & Haynes, 1975; Gordon & Mower, 1980; Miller, Ott, Berk, & Springer, 1974) or assist in generalizing a Pavlovian response to a novel context (e.g., Mower & Gordon, 1983). That reminder treatments improve performance on a later test in animals is very similar to the effects of practice testing learned information in humans. Additionally, the use of such treatments in animal research was used to identify the role of retrieval in models of Pavlovian learning (Balaz, Gutsin, Cacheiro, & Miller, 1983).

Pavlovian preparations which expose animals to conflicting information across different phases of the experiment (e.g., acquisition followed by extinction) often demonstrate a deficit in retention of the second learned information following a long interval (i.e., so-called

spontaneous recovery) or a shift to a new context (i.e., renewal). This has led to the interpretation that second learned conflicting information is more context specific (Bouton, 1993). Using rats in a two-phase proactive interference situation, Wheeler and Miller (2007) demonstrated that providing an immediate test following an extinction treatment can prevent this recency-to-primacy shift which otherwise occurred following a long retention interval or a shift to a test context different from the one used for target training. Apparently, providing this immediate test 'stamped in' the novel information of nonreinforcement, thereby preventing its degradation due to changes in the temporal or physical context. Speaking to this view, attempts in humans have attempted to use testing as a means to explicitly revise or add to an existing set of information. Using tests as a means to reactivate a previously acquired association between a person's face and their name into a more labile state, Finn and Roediger (2013) introduced a new piece of information (e.g., occupation). The suggestion that retrieved memories undergo some additional encoding at the time of retrieval suggests that retrieval might be used to more effectively update or revise stored memories (e.g., Alberini, 2011; Bjork, 1975) than further study. Finn and Roediger anticipated that retrieval processing of the face-name association would allow for better integration of the new occupation information than would additional study of the name-face pair. Surprisingly, they observed that the act of retrieving the name-face association actually impaired acquisition of the novel information. Such retrieval practice has also been shown to prevent interference of the retrieved information (Potts & Shanks, 2012). The retrieved association seems to be dominant over the second-learned association, although in Finn and Roediger's preparation the second-learned information actually followed practice testing. The immediate suggestion is that this finding might illustrate that retrieval of the previously learned association serves to block acquisition of the new association (Kamin, 1969), with retrieval enhancing the association of the blocking association (face-name) at the expense of the novel association (face-occupation). Finn and Roediger suggest that blocking is an insufficient explanation because successful retrieval of the face-name association was positively related to whether the occupation was recalled; however, they admit this correlation could be the product of item selection (e.g., highly salient faces) which would make this an invalid measure of the degree to which blocking plays a role. Additionally, the novel information immediately after testing might not have been provided during the appropriate temporal window that might have facilitated integration of the new material (e.g., Monfils, Cowansage, Klann, & LeDoux, 2009). Finally, it seems that one must ask whether the task demands themselves, retrieval processing of any information, might have prevented processing new information. Although this last interpretation is not an exciting theoretical interpretation, it might be illuminating to include a control for retrieval of irrelevant information, in addition to the study-only control.

The preceding remarks concerning the memorial benefits of retrieval in nonhumans admittedly involve procedures highly dissimilar to those used in studies of the testing effect in humans. Notably, in the above studies with animals, the target information was usually memory of a single event, whereas in studies of the testing effect in humans, there are ordinarily a series of memories. However, there is one experiment with rats in which the target information was a sequence of events. Miller (1982) showed that a sequence of instrumental responses by rats can also benefit from reminder treatments, and

in so doing provided what might be viewed as the first systematic demonstration of the testing effect in nonhumans (although his studies did not demonstrate that retrieval was actually *superior* to an equivalent number of further training trials). In Miller's Experiment 4 rats were given a few training trials on a complex maze using either sucrose (present at the goal box) or continuous foot shock (absent at the goal box) as reinforcement. The critical manipulation was that after three trials in the maze half the animals, orthogonal to reinforcement type, were give brief (30 s) exposure to the start box of the maze in the absence of reinforcement as a form of CS-alone reminder trial for the contextual cues (i.e., retrieval practice). The working assumption is that exposure to the start box reactivated the stored memory of the earlier training trials. Much like their human counterparts, the rats that received this retrieval practice improved their performance on later test trials relative to rats that were not given the reminder trials. Miller's parametric data with amount of training before the nonreinforced reactivation (i.e., testing-like) trials demonstrated that the testing effect was maximal when there was sufficient prior training for there to be a memory to reactivate, and not so much training that performance was maximal which would have masked any potential improvement due to memory reactivation. One can regard this point as defining one of the critical boundary conditions for observing the testing effect. The other important boundary condition is that many cue-alone reactivation trials often yield extinction of the target behavior rather than the enhancement seen to result from a small number of reactivation trials. Of course, with human subjects, instructions can minimize this appearance of extinction with extensive rehearsal practice, as seen when a person is asked to covertly rehearse some target material.

In a similar vein, using young rats Campbell and Jaynes (1966) reported that reinstatement of the cues from training retarded the decay of performance as the retention interval increased. This demonstrates that a series of retrieval trials spaced over a retention interval can attenuate or even prevent 'forgetting' over the retention interval that would otherwise have occurred. Here retrieval is not so much about improving memory in absolute terms, as preventing a loss of memory (or at least performance) in the control group. Specifically, Campbell and Jaynes showed that reinstatement trials reduced infantile amnesia in young rats. Thus, both Pavlovian and instrumental learning appear to benefit in terms of enhanced recall at a later test from treatments that provide the organism with an opportunity to practice retrieval.

Theoretical accounts of the testing effect

Research interest in practice retrieval as a strategy to enhance subsequent test performance in humans has become increasingly popular over the past 30 years. This was primarily due to robust evidence that taking a practice test results in better performance on later tests than does additional exposure to that material (i.e., additional study). Interest in the underlying basis of the testing effect lagged behind observation of the phenomenon itself. Below we review the more popular accounts of the testing effect, and the paradox between the testing effect and associative extinction. As will be seen, no one account suffices to explain all of the reported data. Likely, the testing effect arises from several different mechanisms, with the relative contributions of each mechanism varying across situations (e.g., Yang et al., 2021).

Testing as additional study

A frequently offered explanation of the testing effect is that the additional testing merely serves to provide additional exposure to the learning material. This account is based on the assumption that testing is an opportunity for additional processing of the original information especially when testing includes feedback (Kolers, 1973; Skagg, 1920), thereby either strengthening the [singular] target association or laying down additional copies of the target association (i.e., an instance framework). Early empirical demonstrations of the testing effect which lacked appropriate study-only controls are certainly subject to this interpretation. However, in the last thirty years investigations of the testing effect have clearly demonstrated that testing provides a retention benefit even when the amount of exposure to the tested information is matched in control conditions (e.g., Carrier & Pashler, 1992; Roediger & Karpicke, 2006a). In fact, several demonstrations provided *more* time during additional study than it took to perform the practice test (Kornell et al., 2009). The account of the testing effect asserting that it is no more than additional study has also been rejected on the basis that both longer (but not too long) retention intervals between practice tests and more difficult tests improve later recollection, indicating that the amount of processing is not so much important as the act of retrieving the information and the completeness of those retrievals (Carpenter & Delosh, 2006, see below; Glover, 1989; Pashler et al., 2003). Thus, this account is certainly not viable as a complete account of the testing effect; however, it is important to consider the role of additional study when generating control conditions for future empirical investigations. In this sense, the testing effect may be more conservatively defined as the enhancement due to testing relative to the benefits of additional study.

Although the benefit of testing in humans is observed in many different types of learning, one might try to reduce the phenomenon of the testing effect to simple Pavlovian conditioning, in which initial study parallels basic acquisition of an association between two events. This account is most readily applied to those studies which used paired associates as the content of learning. Presenting one of the events to determine whether it activates a representation of the other is standard practice for assessing the strength of such associations. But presenting a cue without its outcome constitutes an extinction trial and in most theories of learning should reduce the strength of the association. However, if one posits that co-activation of representations of a cue presented in the external environment and an outcome retrieved from reference memory creates a condition that potentially supports strengthening the association (e.g., Konorski, 1967). Additionally, one might turn to reconsolidation accounts which posit that activation of a memory or an associated pair allows for a temporal window wherein that association is particularly labile (e.g., Nader, Schafe, & La Doux, 2000). The problem with such accounts is that they must account for both the effect of retrieval practice and conventional extinction, which is challenging because the operations are the similar (i.e., presentations of the CS alone) while the behavioral consequences are diametrically different. One might view the enhancement in responding constituting the testing effect and subsequent decrement in responding constituting extinction to be a function of the number of nonreinforced trials, with few trials of the cue presented alone along with the retrieved representation of the outcome adding to the excitatory value of the cue, and further presentations of the cue alone

reducing responding to the cue. However, for an alternative to 'number of trials' as the key to when a testing effect as opposed to extinction will be observed, see the section below on 'Implications of the testing effect for models of learning' for an alternate account of the distinction between retrieval practice and extinction.

Practice tests improve learning on subsequent target training trials

Recent evidence lends support to the view that practice tests both increase motivation to learn during subsequent study and/or practice tests, and improve the focus of attention on the appropriate test cues and target material. By increasing motivation and appropriate attention on trials that follow the practice test and precede the final test, new learning on subsequent study trials and/or subsequent retrieval practice trials should be enhanced (for a detailed review of studies supporting this account, see Yang et al., 2021). This facilitation of trial following the practice test is sometimes called the *forward test effect* (as distinct from the conventional testing effect which is sometimes referred to as the *backward testing effect* because it seemingly improves retrieval of previously learned material). The magnitude of this sort of benefit is apt to depend on the specific format of the practice test relative to future study trials (Bjork & Storm, 2011). To the extent that there are target training trials after retrieval practice, the forward testing effect as well as the conventional [backward] testing effect could contribute to the observed enhancement produce by retrieval practice.

Transfer-appropriate processing

The transfer-appropriate processing account of the testing effect is that the enhanced performance arises from the similarity between how information is encoded during practice tests and how information is retrieved during the final test (McDaniel & Fisher, 1991; Morris et al., 1977; but see Surprenant & Neath, 2009, for a discussion of how processing that helps one discriminate between correct and incorrect information, rather than similarity, is critical for transfer-appropriate processing). The design for examining the testing effect typically uses practice tests that are at least similar, if not identical to the test used to measure final performance. The transfer-appropriate processing account of the testing effect in lay language is sometimes described as 'practice at retrieval' improving subsequent retrieval. One prediction of the transfer-appropriate processing account is that if a practice test (e.g., free recall) differed from the final assessment (e.g., multiple choice), then the testing effect should be attenuated compared to the practice and final being identical. In some situations in which different tests were implemented for practice and final assessment, the benefit of practice tests was appreciably reduced (Johnson & Mayer, 2009; McDaniel & Fisher). These findings suggest that at least one reason practice tests are found to provide a benefit over further study of the material is that testing involves similar processing (i.e., retrieval) during both practice tests and final assessment.

Retrieval practice may provide a benefit over further study, however, even when assessments vary. The emphasis of the transfer-appropriate processing account of the testing effect is on the overall similarity between tests rather than the retrieval process itself, although requiring retrieval is one way in which all tests are to varying degrees similar. Granted, the evidence that the testing effect does sometimes transfer across different types of tests implies that the transfer-appropriate processing account is limited (Butler, 2010). However, this is correct

only if one takes the rather unrealistic approach that the underlying information processing has no similarity across tests of different types. What is needed to claim that the benefits of testing are due to a match between practiced retrieval and final retrieval is a situation in which retrieval practice provides more of a mismatch in processing during final assessment than does the processing of a study-only control. Given that such a situation is difficult to conceptualize, this possibility has yet to be examined. Final assessment will always require some kind of retrieval process; therefore, one can argue that the retrieval process, in general, is the key similarity that allows for generalization across differing types of testing. However, a negative testing effect has sometimes been observed even when there is a strong match between retrieval and practice tests (e.g., when the test requires comparably superficial processing; Halamish & Bjork, 2011). This finding is interpreted as support for the asymmetric benefits that are offered by retrieval practice leading to a bifurcated distribution of memory strength. Content which was not successfully retrieval practiced is more likely to be missed on the subsequent test than is studied information. This account is discussed further below in the section on Bifurcated distribution.

Returning to Glover (1989), we see here a well-designed assessment of the transfer-appropriate processing account of the testing effect that compared the magnitude of the testing effect following different types of practice tests as a function of the nature of the final test. In Experiment 4a-c of Glover's studies, participants were asked to first read an essay and instructed to return for Phase 2 after a two-day delay. During Phase 2, participants were placed into one of four practice test conditions, free-recall, cued recall, recognition, or no test. Participants were then given a final test four days later. In Experiment 4a, the final test consisted of a free recall test, whereas Experiment 4b used cued-recall as the final test and Experiment 4c used recognition as the final test. Looking across these studies, a transfer-appropriate processing account would expect that performance would be better on the final test when the final test matched the practice test. Inconsistent with the transfer-appropriate processing account, the free recall practice tests produced reliably better performance in each of the final assessments. These findings suggest that it is the completeness of what had to be retrieved during practice testing that provides a benefit to later recollection of the tested material (an alternative account in terms of effortful retrieval is considered later). Retrieval during a free recall test is more complete than during cued recall or recognition due to the fact that the cue present during cued recall and recognition provide a fraction (or all) of the target material, so that less retrieval is required. However, Yang et al. (2021) clearly demonstrate that testing effects can be obtained with mismatched tests; they simply tend to be more pronounced when the tests match.

Carpenter & Delosh (2006) successfully replicated and extended the findings of Glover (1989), correcting the initial failure to equate exposure during the practice test phase. Additionally, they found that a cued-recall practice test was superior to a recognition test, thus providing additional evidence that the completeness what is recalled is what is critical for the testing effect. Carpenter & Delosh went on to argue that retrieval practice produces a benefit by forcing the participant to a make use of a more elaborate retrieval route (more complete retrieval presumably requires more elaborate routes). This conceptualization of the testing effect as producing multiple retrieval routes assumes that the more strategies one has for retrieving the information, the more likely it is that overall retrieval will be

successful (Bjork, 1975, 1988; Estes, 1955; McDaniel & Masson, 1985). Support for this account stems from findings in which variability in the test type when practicing retrieval improves retention (McDaniel & Masson; but see Butler, 2010). Carpenter and Delosh's (2006) Experiments 2 and 3 used partial word stems as a cued-recall task in order to manipulate the number of letters in the test cue. Fewer letters in the word stem increases the number of possible solutions and encourages more elaboration. The argument that fewer cue letters require more elaborate processing was based on the assumption that a larger portion of the lexical neighborhood must be navigated in order for the participant to reach a solution. This assumption ignores that possibility that fewer retrieval cues also will increase the difficulty of retrievals (see below).

Effortful retrieval

The amount of 'effort' that goes into learning has long been known to positively correlate with retention of the material (Craik & Tulving, 1975; Kane & Anderson, 1978). This observation has been elaborated into a more general framework by Bjork and Bjork (1992) in their *new theory of disuse* which emphasizes retrieval strength as a 'use it or lose it' attribute of any given association. The new theory of disuse, which consists of an empirically based set of assumptions about human memory, has proven to be a useful tool for generating novel predictions regarding retrieval practice. Within this framework, there is a distinction between *storage strength* for an item and the *retrieval strength* of that item. Storage strength is defined as how well encoded the item is, whereas retrieval strength describes the ease of access to the item. Storage strength of an item is assumed to never decrease, and storage capacity is unlimited. Changes in storage strength are always positive, and negatively accelerated with increasing numbers of training trials, indicating that there are diminishing returns for storage strength from each additional trial. Retrieval strength across items is assumed to be limited both in the quantity of items that can be retrieved at a single moment and the total retrieval strength of those items. In other words, retrieval strength is unable to increase for one item without decreasing the retrieval strength of other items. Changes in retrieval strength can either be positive or negative and any increments or decrements are assumed to decelerate as they approach asymptotic values (e.g., zero and one). The existing storage strength for an item is assumed to modulate changes in retrieval strength for that item, such that strong storage strength enhances any increments to and reduces decrements to retrieval strength. The final and most critical assumption for our present purposes is that the act of retrieving an item from storage provides larger increments to both storage and retrieval strengths than when items are merely studied and the more difficult the retrieval is, the larger the benefit. Thus, the new theory of disuse anticipates the testing effect not because it provides a mechanism but because it simply incorporates the effect as this last assumption. Additionally, changes in storage and retrieval strengths on test trials are assumed to not occur without successful retrieval, which congruent with the basic premise of *desirable difficulty* that retrieval is optimally beneficial when it is effortful but successful (Bjork, 1994).

The new theory of disuse provides a qualitative account of many findings with respect to the testing effect. However, other than a description of how retrieval strength and storage strength change with each successful retrieval trial and the positing that both

storage and retrieval strengths are necessary for recall, this theory lacks an underlying mechanism for many of its predictions. Given that the theory is vague, it lends itself to being roughly applied to many circumstances. If Bjork and Bjork (1992) had proposed a specific underlying mechanism for their assumptions, the model would make more detailed predictions which would make testing it easier. For now, the theory of disuse is a relatively safe theory in that it's basic premise (i.e., retrieval difficulty on a test improves later retention) seems to be valid; however, the model does not provide sufficient detail to go much beyond that prediction. To illustrate, consider Karpick and Roediger's (2007) study comparing equally spaced versus expandingly spaced retrieval practice trials. They replicated the findings of Landauer and Bjork (1978), that expanding retrieval practice produced better performance after a short delay, but equally spaced retrieval practice produced better retention following a long delay. In their critical experiment (Experiment 3), they observed that a long delay between study and retrieval practice produced the same improved long-term retention independent of whether successive retrieval trials were expanding or equally spaced. The new theory of disuse can account for both findings because it does not commit to whether initial retrieval difficulty or later retrieval difficulty is more desirable. Predictions at this level would be clearer if the model were mathematically formalized.

Pyc and Rawson (2009) provide compelling evidence that the testing effect is well described in terms of retrieval difficulty. This was accomplished by manipulating retrieval difficulty using two different strategies, increasing the retention interval between successive retrieval practices (i.e., expandingly spaced trials) and the number of times each item had to be retrieved. Increasing the retention interval has been manipulated in a number of different ways and it has been consistently found that retrieval is more difficult as intervals expand (e.g., see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). Manipulating the number of successful retrievals appears less frequently in the literature, but the theoretical justification provided by the new theory of disuse is that with each successful retrieval, future retrievals become easier. Therefore, the number of retrieval trials is thought to facilitate retrieval in addition to any additional increase in storage strength with the additional trials. The assumption that each of these manipulations alters difficulty of retrieval has empirical support (Karpicke & Roediger, 2007; Bjork & Bjork, 1992). In order to test the predictions regarding the effect of retrieval difficulty as captured by intertrial interval and number of retrievals, Pyc and Rawson presented participants with an initial study phase of 70 foreign-English paired associates. These 70 pairs were used to create 10 different study lists. Seven of these lists were assigned various criterion levels for practice testing, which determined how many successful retrievals (1, 3, 5, 6, 7, 8, or 10) were required before the item was dropped from the remainder of the session. The number of successful retrievals was manipulated within-subjects. In order to manipulate the interval between retrievals, some participants practiced retrieving the English paired associates with lists of six foreign items (short interval condition), whereas other participants were tested with lists containing 30 items (long interval condition). Participants in both conditions continued retrieval practice on their respective list until they reached criterion performance for each item. There was no explicit feedback to these participants, and the within-subject criterion manipulation presumably minimized participants' ability to rely on implicit feedback as certain word pairs

appeared more frequently or less frequently during retrieval practice. Thus, participants in the short interval condition received practice tests on a total of 10 different lists, whereas participants in the long interval condition received practice tests on only two lists. Consequently, each group of participants experienced the same number of total retrievals over the same amount of time; however, shorter lists ensured that repeated items would occur closer together in time. Their findings indicated that long intervals between trials produced better recollection on the final test. Additionally, more retrieval trials were found to improve recollection. But their critical finding was that successive retrievals provided diminishing gains in performance. The data indicate that the benefits provided by additional retrieval was nonlinear, suggesting that successive retrieval practice provided less of a benefit, presumably because those retrievals were easier retrievals.

In some sense the testing effect may be related to the generation effect, in that both rely on the participant retrieving a response from memory; however, there is a notable distinction. The generation effect occurs when participants are given a partial word stem and asked to fill in the missing letters (for a review, see Bertsch, Pesta, Wiscott, & McDaniel, 2007). When participants generate the word themselves, their retention for that word is better than if they were simply asked to read the word. The generation effect occurs when retrieval of the word is incidental to filling in the missing letters, whereas in the testing effect the participant is explicitly trying to recall the target word. While words recalled in both situations may be identical from the perspective of the participant, the retrieval mode is different. Karpicke and Zaromb (2009) compared the testing effect to the generation effect by simply manipulating the instructions. After an initial study of word lists, participants were provided with partial word stems and instructed to either fill in the blanks with the first word they could think of (generation effect) or to use the word stem as a retrieval cue to recall one of the words from the previous list (testing effect). The results of their four experiments suggested that generation and testing effects produce differential results. Experiment 1 clearly demonstrated a retrieval benefit provided by the explicit retrieval condition; however, a generation effect was not observed. In Experiment 2, when the generation effect did produce better recall than the read-only condition, the benefit of the retrieval test condition over the generation condition was only marginal. Experiments 3 and 4 investigated whether the instruction manipulation produced a difference on a final recognition test and found a clear benefit of explicit retrieval on recognition (Experiment 4 only).

The generation effect and testing effect involve retrieving a particular item based on the same set of retrieval cues. The only difference is that during the practice period the participants in the testing condition are instructed to retrieve items only from the list previously studied, whereas generation involves the participant selecting from their entire vocabulary to fit the present constraints (e.g., word stem, sentence completion, etc.). This is analogous to the influence of additional lures in a multiple-choice task which seemingly makes it more likely to select a potential distracter. In practice, it has been found that adding lures to a multiple-choice practice test can actually increase final performance as long as many errors are not committed during practice (Butler, Marsh, Goode, & Roediger, 2006). The finding of Butler et al. fits well with the concept of desirable difficulty because increasing the number of lures should add to the difficulty of the task. However, if the

retrieval is too difficult and successful retrieval does not occur, then there should be no benefit of retrieval. If we can equate the generation condition of Karpicke and Zaromb (2010) with increasing the number of lures, then a retrieval difficulty account would predict more of a retrieval benefit in the generation condition. Karpicke and Zaromb recorded the number of successful completions of word stems in both the generation and retrieval practice conditions and found similar levels of performance initially. However, performance on a final recall test was improved in the retrieval practice condition relative to generation. One might explain this discrepancy by pointing out that in the retrieval practice condition learning was explicit. Consequently, the participants may have been rehearsing the word after successful retrieval, whereas in the generation condition once the word was generated, participants likely were not actively trying to retain the information.

The general framework of retrieval difficulty, as suggested in Bjork & Bjork's (1992) new theory of disuse, provides a flexible qualitative model for the testing effect. Additionally, this approach has generated a number of research questions, the answers of which clarify the specific conditions under which testing will provide a benefit and informs us how to maximize that benefit (e.g., Cepeda et al., 2008). However, something that this account must struggle with is that Yang et al. (2021), in a detailed meta-analysis, found that recognition tests produce greater testing effects than does free recall.

Retrospective revaluation accounts

Retrospective revaluation refers to a change in the behavior reflecting a target cue-outcome association as a result of altering the associative status of nontarget cue that was present when the target cue had been paired with the outcome (e.g., backward blocking). Some instances of the testing effect lend themselves to be explained by models that are successful accounting for retrospective revaluation (Dickinson & Burke, 1996; Stout & Miller, 2007; Van Hamme & Wasserman, 1994). Consider situations in which human subjects are asked to covertly rehearse material that they were previously trained on, with neither the target cues or outcomes themselves actually being presented, or Miller's (1982) previously described maze studies with rats. In the latter case, retrieval practice was initiated not by presentation of the target CSs (i.e., choice points in the maze), but by the start or goal box of the maze which presumably reactivated memories of both the choice points and the reinforcer. The activation on these retrieval practice trials of both the cue (or response in Miller's instrumental task) and outcome by another stimulus that was present during initial training create the conjoint absent-cue and absent-outcome activation condition that the accounts of retrospective revaluation predict would increase behavior indicative of the target association. However, the retrospective revaluation accounts are unable to explain the testing effect when the retrieval practice trials include presentation of the target cue without the target outcome, which is what in fact is done in many demonstrations of the testing effect (Roediger & Karpicke, 2006a). Thus, the retrospective revaluation accounts at best succeed in explaining only the sub-class of testing effect in which neither the target cue nor target outcome is presented during retrieval practice. Note that the preceding discussion assumes that (a) the enhanced performance presumably resulting from the target cue causing retrieval of the outcome representation, and (b) the enhanced performance presumably resulting from retrieval of the target cue and target outcome representation by presentation of a common

associate are due to the same underlying process(es). The validity of this assumption should be addressed in future research.

Bifurcated distribution

Another framework for conceptualizing the benefits of testing relative to additional study is to think of the target content as being composed of many items, each with its own memory strength. Additional study serves to increase the memory strengths of all items to some degree, whereas practice testing improves the memory strengths of retrieved items by a larger amount, but leaves the memory strengths of those items which were unsuccessfully retrieved relatively unchanged (Halamish & Bjork 2011; Kornell, Bjork, & Garcia, 2011). Thus, the content is split, or bifurcated, into items with robust retrieval strength (after a successful retrieval) and relatively low retrieval strength (after a failed or interfering retrieval) following practice testing. The benefit of testing over that of additional study is assumed to be the increased memory strength to these select items that were successfully retrieved during practice testing. This model is useful for understanding the standard influence of delay on the testing effect. When testing is immediate, performance might be best when all items have received a minor boost in retrieval strength. As testing is delayed, the threshold for retrieving an item increases, meaning that more items from the restudy condition will appear to have been forgotten relative to the items that were successfully retrieval practiced. One would expect to see this pattern as a function of retrieval difficulty, not merely increasing delay, as demonstrated by Halamish & Bjork when they observed the reverse testing effect with an easy (cued recall test) relative to a difficult (free recall test). Halamish and Bjork also support their claim that retrieval benefits are more prevalent when testing is difficult by introducing associative interference to make the cued recall test more difficult.

One merit of the bifurcated distribution account is that it gets away from questionable assumptions like transfer-appropriate processing to account for retrieval-practice effects and offers a rich account in terms of retrieval difficulty, but stumbles on its reliance on the vague concept of memory strength (Roediger, 2008). With the bifurcated account, Bjork and colleagues are expanding on their prior concept of ‘desirable difficulty’ during practice testing, but now adding that retrieval difficulty of the final assessment also determines whether the mnemonic benefits of practice testing will emerge. One method of preventing some items from lagging behind is to provide corrective feedback during retrieval practice, which at least allows those items to gain some retrieval strength and increases their potential to be recalled on subsequent tests. Notably, Storm, Friedman, Murayama, and Bjork (2014) found that, although practice testing can enable a benefit on a delayed (1 week) final test, if feedback is given on that test, one can see that benefit vanishes and instead previously studied items are better recalled, that is, a reversed testing effect is observed. This finding fits with the assumption that initial retrieval practice without feedback resulted in a bifurcated distribution of retrieval strengths for the items with some being very strong and some very weak, whereas additional study produced a distribution of retrieval strengths that was moderate. The test with feedback presumably increased the putative retrieval strength of all items, presumably allowing the studied items to begin to exceed the threshold for retrieval whereas forgotten retrieval practiced items where still too weak to be recalled on

the subsequent test. Storm et al. effectively demonstrated that these relative differences in retrieval strength persist over long delays and can be illustrated by giving all the items additional practice rather than directly manipulating the test difficulty.

Episodic context account

Karpicke, Lehman, and Aue (2014) offer an account of the testing effect in terms of temporal contexts which serve as reminder cues for retrieval practiced information. They suggest that information is encoded during initial study along with the episodic context of that event. Over time, context, particularly the temporal components of context, are assumed to change and become increasingly different from that initial study context. When information is successfully retrieved, the contextual information at the time of retrieval becomes integrated with that retrieved information. This effectively allows retrieval practice to function as training the target information in multiple contexts, which has been shown to help that information generalize to novel contexts. Merely studying the information again does not seem to as effectively update the contextual information linked to the target material because this process is thought to happen during retrieval. Some retrieval may occur during study; however, these incidental retrievals are thought to be less potent and less frequent than when the task specifically calls for retrieval of the target material. Finally, subjects are thought to use contextual information to aid retrieval on subsequent tests. Having been associated with a larger variety of contexts, the previously retrieval practiced information has more retrieval cues present at these later tests. This account is compatible with the retrieval difficulty account because practice tests that are difficult are assumed to have subjects recall the target material with fewer retrieval cues. This more difficult retrieval is thought to strengthen later retrieval of the target material. The novel aspect of the episodic account is that it offers an explanation of Karpicke and Zaromb (2010) finding that retrieval practice enhances future retrieval more so than does generation. In their retrieval practice condition, subjects were encouraged by the instructions to use the contextual information from the initial study to aid retrieval during retrieval practice. This presumably facilitated retrieval on the final test by enhancing the effectiveness as retrieval cues of the contextual cues that had been present during initial study.

Post-retrieval monitoring

The previous accounts were largely couched in terms of improving retrieval of the target material on the final test. An alternative view is that practice testing increases monitoring of retrieved material on the final test. Pierce et al. (2017) reported a clever series of experiments suggesting that at least in some demonstrations of the testing effect the critical underlying process was that as a result of practice tests the participant was better able on the final test to discriminate between retrieved target material and retrieved nontarget material. Further research is needed to determine the generality of this account of the test effect.

Implications of the testing effect for models of learning

The testing effect provides a theoretical challenge to most contemporary associative accounts of learning, which largely ignore the potential for presentation of only one component of an association to improve subsequent performance dependent upon retrieval

of the complete association. Whereas the testing effect with feedback might be viewed as a form of additional cue-outcome training (but see Hays et al., 2010), the testing effect without feedback presents a special challenge if applied in an equally reductive manner given that it most closely resembles an extinction trial that would typically undermine the original training. However, the reliable and robust observation of an enhancement of performance following testing even without feedback that constitutes the testing effect appears to arise from additional processing of the previously acquired information that is not captured in any current associative account. This failure is primarily the result of an emphasis on trial-wise models of learning that focus on degrees of contingency and contiguity between the paired events during training, with little-to-nothing being said concerning processing at the time of testing (e.g., McLaren & Mackintosh, 2002; Pearce & Hall, 1980; Pearce, 1987; Rescorla & Wagner, 1972; but see, Stout & Miller, 2007; Wagner, 1981). Such perspectives are challenged by the prospect that presenting only a cue, in the absence of the outcome, could improve later recall of the cue-outcome association. For example, the Rescorla and Wagner (1972) model anticipates that presentation of a cue on a test trial should activate a representation of the outcome (i.e., an expectation of the outcome) that is then not presented, a procedure that constitutes an extinction trial. According to that model, the surprising absence of the outcome based on presentation of the cue determines the amount by which the strength of the association should be reduced (i.e., extinction). This is clearly in opposition to the widely observed testing effect.

A CS presented alone following CS-US acquisition trials in a Pavlovian preparation closely resembles an extinction trial. An extinction trial should be detrimental to expression of the previously learned CS-US relationship, and repeated extinction trials are in fact ordinarily detrimental for these predictive relationships (i.e., Pavlov, 1927). However, when humans are placed in settings like the preparations reviewed previously, it is generally ‘understood’ by the subject that the retrieval practice phase is distinct from the acquisition phase and is perceived as assessment of the acquired information, or a covert rehearsal trial if responding is not requested, as distinct from another training trial, whereas in Pavlovian tasks (with nonhuman animals and sometimes humans) the retrieval practice phase is not distinct from acquisition and is presented as a new training contingency that directly conflicts with the previously learned contingency.

One potential way to differentiate between a [practice] test trial and a typical extinction trial in Pavlovian preparations would be to vary whether the outcome is explicitly absent (i.e., an extinction trial), or a situation in which the absence of the outcome is ambiguous. An ambiguous or unknown outcome presentation on a rehearsal trial on which only the cue is presented may reduce or eliminate the amount of conflicting information learned on that trial, while still potentially providing both retrieval practice and, if responding is allowed, assessment of the retrieved association. Blaisdell, Leising, Stahlman, and Waldmann (2009) demonstrated this in a sensory preconditioning preparation with rats. An auditory cue was paired with presentation of a light, followed by the light being paired with food. At test, the auditory cue elicited food seeking behavior in control subjects, but that behavior was reduced if prior to testing the auditory cue was repeatedly presented without the light being turned on. However, if the light bulb was covered so the rat could not see whether or not it came on, making extinction of the auditory cue ambiguous, the same

pre-test presentations of the auditory cue did not reduce later behavioral control by the auditory cue (see Waldmann, Schmid, Wong, & Blaisdell, 2012, for similar findings in first-order conditioning). Similar manipulations have been undertaken to investigate differences in learning about implicitly absent events relative to explicitly absent events in human contingency learning (Castro, Wasserman, & Matute, 2009; Wasserman & Castro, 2005). The findings in these studies corroborate the suggestion by Blaisdell et al., that the absence of an outcome that is made explicit is more likely to reverse or interfere with previously learned contingencies. Perhaps paradigms that avoid the explicit absence of outcomes in Pavlovian preparations would provide a better parallel to testing effects with humans because they mitigate against the acquisition of new contingencies which could result in a 'negative testing effect.' Unfortunately, the potential benefits (as opposed to the decrements reported by Blaisdell et al.) for later performance of testing rats when the outcome is obscured have not been adequately assessed, nor has the potential for a testing effect in contingency learning been fully assessed in humans.

The preceding discussion of operations that might differentiate when cue alone presentations are apt to result in a testing effect as opposed to extinction are based on the empirical results of a relatively small number of studies. The proposed importance of the expectation or lack of expectation of overt reinforcement is not anything that is *a priori* anticipated by any of the formalized contemporary theories of associative learning. There is work to be done here connecting these observations to theory. Another possible avenue of research might involve the number of cue alone presentations.

Concluding remarks

In this review, the historical evolution of the testing effect has been summarized from initial empirical findings to more recent investigations of the effects of practice tests on subsequent performance. Much of the emphasis in prior discussions of the testing effect has been on the applied value of testing as a tool to be implemented in educational settings. Here we examined topics including the role of intervals between initial study and retrieval practice, intervals between successive retrieval practices, and intervals between retrieval practice and final testing in determining when final retrieval will be enhanced, as well as generalization of retrieval practice across different retrieval tasks, the role of that feedback plays in the testing effect, and the seemingly inconsistency between the testing effect and extinction phenomena. Finally, summary descriptions of several of the proposed accounts of the testing effect were presented and evaluated. In doing so, it became clear that the testing effect is not due to mere additional exposure to the target material as was initially proposed. The benefit of testing is clear only when appropriate controls for exposure alone are used to compare with the effects of testing per se. Additionally, although transfer-appropriate processing provides a reasonable account of the testing effect under some circumstances, there is accumulating evidence that tests which require more effortful processing (e.g., free-recall) provide more of a benefit on an easier final test (e.g., cued-recall) than if the practice test perfectly matched the final assessment (Carpenter & DeLosh, 2006; Glover, 1989; see also Carpenter, 2009). Thus, it seems that a more general process such as surmounting retrieval difficulty better captures the benefit offered by testing. This suggests that although transfer-

appropriate processing may influence the testing effect, it is not a complete explanation of the phenomenon.

It is worth noting that just as cue-alone presentations after training can both enhance subsequent conditioned responding (i.e., the testing effect) and impair subsequent conditioned responding (i.e., extinction), so too does outcome-alone presentations. When the outcome is of biological relevance (i.e., a US), enhanced subsequent responding to the cue is commonly observed, a phenomenon often called *reinstatement* (Rescorla & Heth, 1975), which is often explained in terms of heightened motivation by strengthening the representation of the US, or, given its context dependency, summation of US activation by the CS and the reinforced context. But facilitated subsequent retrieval similar to that proposed for the testing effect may play a role. Balaz et al. (1983) spoke about reinstatement effects as arising from a US-alone presentation serving as a super salient retrieval cue that engenders successful retrieval of the target association when an ordinarily retrieval cue (e.g., the CS in a Pavlovian preparation) proves to be an inadequate retrieval cue. Although enhanced subsequent responding is often seen, like cue-alone presentations, impaired subsequent responding is sometimes reported, and when it is, the effect is often explained in terms of degraded contingency. Thus, we see here [another] seeming symmetry between the processing of CSs and USs (Gunther, Miller, & Matute, 1997).

Accounts of the testing effect are offered by a broad set of theories that suggest that testing increases the number of retrieval routes to access the target information and/or that surmounting the difficulty posed by the retrieval process during retrieval practice facilitates subsequent retrieval. However, the relationship, if any, between retrieval difficulty and the process of increasing retrieval routes is not well understood. In the literature, these two approaches to the testing effect exist in parallel, accounting for similar findings using slightly different terminology. That is, tests that encourage multiple retrieval routes may simply be more difficult, and more difficult tests may force the development of new ways of retrieving the relevant information. The major distinction between these two, not necessarily mutually exclusive, accounts lies in the actual mechanism involved during retrieval practice. Does practicing retrieval force the learner to access the information in novel ways? Or is it simply that surmounting the difficulty in retrieving the information that improves subsequent retrieval by better engaging the same retrieval processes that were engaged on the practice test? These questions provide a difficult challenge to future researchers because the questions require differentiating two processes that may be intrinsically related. One benefit offered by the effortful processing account is that, within a given task, it is often possible to vary the difficulty in systematic ways (e.g., Finley, Benjamin, Hays, Bjork, & Kornell, 2011; Pyc & Rawson, 2009). The difficulty of a task is more easily operationally defined than is the concept of flexible retrieval routes, but both suggestions could benefit from being more concretely defined.

The benefit with respect to future retrieval provided by practice tests without feedback (e.g., Butler & Roediger, 2008; Miller, 1982; Potts & Shanks, 2014) is fundamentally problematic for associative theories of learning. Although testing effects have been observed in both associative and instrumental preparations, associative theories (e.g., Rescorla & Wagner, 1972) consistently predict that CS-alone trials (i.e., practice tests) following initial

learning (i.e., CS-US pairings) constitute extinction trials and should be detrimental to the responding based on the initially learned CS-US relationship. Perhaps the definition of extinction needs to be amended considering the distinction between explicit absence and ambiguity (Blaisdell, Leising, Stahlman & Waldman, 2009; Castro et al., 2009; also see Fast & Blaisdell, 2011). When feedback is provided (i.e., US presentation in Pavlovian situations), extinction is not expected because this makes the practice test more like an additional acquisition trial. Although feedback enhances the testing effect in many circumstances, feedback is not always necessary to observe a benefit from practice tests (Pashler, 2005). Bjork and Bjork's (1992) 'new theory of disuse' attempts to address this failing of associative models by differentiating 'storage strength' from 'retrieval strength;' however, the model still lacks formalization. The concept of retrieval practice in general is not compatible with contemporary models of associative conditioning because most of these models represent learning purely in terms of storage (i.e., associative) strength. Even models that are retrieval focused (e.g., Miller & Matzel, 1988) represent associative strength in terms of multiple storage strengths that simply interact at test, and they do not account for benefits of retrieval practice. Therefore, the testing effect presents a challenge for associative accounts that have for too long, ignored the role of retrieval processes. Extensions of the R-W model that are used to address retrospective revaluation through activation and further processing of representations of absent events (e.g., Van Hamm & Wasserman, 1994) are candidates to address this phenomenon; however, further modification would be necessary to capture the observation that cue-specific retrieved information is selectively strengthened even in the absence of external feedback. That is, mere activation may not qualify that retrieved information to be a candidate for improved subsequent retrieval. Perhaps only the strongest activated or a cluster of strongly activated candidates are strengthened. An additional, albeit clunky mechanism might be to assume a sort of internal feedback mechanism that serves to strengthen a cue-outcome association when a cue succeeds at activating a memory of the outcome. This mechanism would be in competition with traditional extinction, but might be appropriately scaled to create an initial boost in the accessibility of that association before new extinction learning becomes dominant with more cue-alone presentations.

Although the testing effect surely has its boundary conditions as all cognitive phenomena do, the testing effect is both robust and pervasive. In light of these two characteristics, perhaps the testing effect should be viewed as a useful benchmark for assessing associative models, or at least a phenomenon we should be designing new models to account for. Given the influence of instruction within so much of the testing effect literature based on human participants, there is a challenge to rigorously examining the it within the framework of basic learning as studied in nonhuman subjects. Immediately, researchers must first explore parameters that may generate a testing effect in more animal preparations or with humans without relying so heavily on instruction or participants' familiarity with test formats.

Acknowledgments

Preparation of this paper was supported in part by NIMH award 033881.

References

Abbott EE (1909). On the analysis of the factors of recall in the learning process. Psychological Monographs, 11, 159–177. DOI:10.1037/h0093018

Alberini CM (2011). The role of reconsolidation and the dynamic process of long-term memory formation and storage. Frontiers in behavioral neuroscience,5, 12–21. DOI: 10.3389/fnbeh.2011.00012 [PubMed: 21436877]

Balaz MA, Gutsin P, Cacheiro H, & Miller RR (1983). Blocking as a retrieval failure: Reactivation of associations to a blocked stimulus. Quarterly Journal of Experimental Psychology, 34B, 99–113. DOI: 10.1080/14640748208400879

Bangert-Drowns RL, Kulik CL, Kulik JA, & Morgan MT (1991). The instructional effect of feedback in test-like events. Review of Educational Research, 61, 213–237. DOI: 10.3102/00346543061002213

Berger SA, Hall LK, & Bahrick HP (1999). Stabilizing access to marginal and submarginal knowledge. Journal of Experimental Psychology: Applied, 5, 438–447. DOI: 10.1037/1076-898X.5.4.438

Bertsch S, Pesta BJ, Wiscott R, & McDaniel MA (2007). The generation effect: A meta-analytic review. Memory & Cognition, 35, 201–210. DOI: 10.3758/BF03193441 [PubMed: 17645161]

Bjork RA (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In Solso RL (Ed.), Information processing and cognition: The Loyola symposium (pp. 123–144). DOI: 10.1.1.694.8821

Bjork RA (1988). Retrieval practice and the maintenance of knowledge. In Gruneberg MM, Morris PE, & Sykes RN (Eds.), Practical aspects of memory: Current research and issues (Vol. 1, pp. 396–401). New York: Wiley. DOI: 1988–97682-062

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

Bjork RA, & Bjork EL (1992). A new theory of disuse and an old theory of stimulus fluctuation. In Healy AF, Kosslyn SM, & Shiffrin RM (Eds.), Essays in Honor of William K. Estes: Vol. 2. From learning processes to cognitive processes (pp. 35–67). Hillsdale, NJ: Erlbaum.

Bjork RA, & Storm BC (2011). Retrieval experience as a modifier of future encoding: Another testing effect. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37, 1113–1124. DOI: 10.1037/a0023549 [PubMed: 21574746]

Blaisdell AP, Leising KJ, Stahlman DW, Waldmann MR (2009). Rats distinguish between absence of events and lack of information in sensory preconditioning. International Journal of Comparative Psychology, 22, 1–18. DOI: 2009–18152-001

Bouton ME (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. Psychological Bulletin, 114, 80–99. DOI: 10.1037/0033-2909.114.1.80 [PubMed: 8346330]

Butler AC (2010). Repeated testing produces superior transfer of learning relative to repeated studying. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36, 1118–1133. DOI: 10.1037/a0019902 [PubMed: 20804289]

Butler AC, Marsh EJ, Goode MK, & Roediger HL (2006). When additional multiple-choice lures aid versus hinder later memory. Cognitive Psychology, 20(7), 941–956. doi.org/10.1002/acp.1239

Butler AC, & Roediger HL (2007). Testing improves long-term retention in a simulated classroom setting. European Journal of Cognitive Psychology, 19, 514–527. DOI: 10.1080/09541440701326097

Butler AC, Roediger HL (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. Memory & Cognition 36, 604–616. doi.org/10.3758/MC.36.3.604 [PubMed: 18491500]

Campbell BA, & Jaynes J (1966). Reinstatement. Psychological Review, 73, 478–480. DOI: 10.1037/h0023679 [PubMed: 5976739]

Carpenter SK (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. DOI: 10.1037/a0017021 [PubMed: 19857026]

Carpenter SK, & DeLosh EL (2006). Impoverished cue support enhances subsequent retention: support for the elaborative retrieval explanation of the testing effect. Memory & Cognition, 34, 268–276. DOI: 10.3758/BF03193405 [PubMed: 16752591]

Carpenter SK, Pashler H, & Cepeda NJ (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. Applied Cognitive Psychology, 23, 760–771. DOI: 10.1002/acp.1507

Carpenter SK, Pashler H, & Vul E (2006). What types of learning are enhanced by a cued recall test? Psychonomic Bulletin & Review, 13, 826–830. DOI: 10.3758/BF03194004 [PubMed: 17328380]

Carrier M, & Pashler H (1992). The influence of retrieval on retention. Memory & Cognition, 20, 633–642. DOI: 10.3758/BF03202713 [PubMed: 1435266]

Castro L, Wasserman EA, & Matute H (2009). Learning about absent events in human contingency judgments. In Watanabe S, Blaisdell AP, Huber L, & Young A (Eds.). Rational Animals, Irrational Humans (pp. 83–99). Tokyo: Keio University.

Cepeda NJ, Vul E, Rohrer D, Wixted JT, & Pashler H (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. Psychological Science, 19, 1095–1102. DOI: 10.1111/j.1467-9280.2008.02209.x [PubMed: 19076480]

Cepeda NJ, Pashler H, Vul E, Wixted JT, & Rohrer D (2006). Distributed practice in verbal recall tests: A review and quantitative synthesis. Psychological Bulletin, 132, 354–380.Bulletin, 144(11), 1111–1146. doi.org/10.1037/bul0000166 [PubMed: 16719566]

Chan JCK, Meissner CA, & Davis SD (2018). Retrieval potentiates new learning: A theoretical and meta-analytic review. Psychological Bulletin, 144(11), 1111–1146. doi.org/10.1037/bul0000166 [PubMed: 30265011]

Chan JCK, Thomas AK, & Bulevich JB (2009). Recalling a witnessed event increases eyewitness suggestibility: The reversed testing effect. Psychological Science, 20, 66–73. DOI: 10.1111/j.1467-9280.2008.02245.x [PubMed: 19037905]

Clarke JC, Westbrook RF, & Irwin J (1979). Potentiation instead of overshadowing in the pigeon. Behavioral & Neural Biology, 25(1), 18–29. doi.org/10.1016/S0163-1047(79)90705-2 [PubMed: 454337]

Congleton A, & Rajaram S (2012). The origin of the interaction between learning method and delay in the testing effect: The roles of processing and conceptual retrieval organization. Memory & Cognition, 40, 528–539. DOI: 10.3758/s13421-011-0168-y [PubMed: 22160872]

Craik FIM, & Tulving E (1975). Depth of processing and the retention of words in episodic memory. Journal of Experimental Psychology: General, 104, 268–294. DOI: 10.1037/0096-3445.104.3.268

Craik FI, & Lockhart RS (1972). Levels of processing: A framework for memory research. Journal of verbal learning and verbal behavior, 11(6), 671–684. DOI: 10.1016/S0022-5371(72)80001-X

Cunningham D, & Anderson RC (1968). Effects of practice time within prompting and confirmation presentation procedures on paired associate learning. Journal of Verbal Learning & Verbal Behavior, 7, 613–616.DOI: 10.1016/S0022-5371(68)80115-X

DeVietti TL, & Haynes DA (1975). Reminder: Similar and differential effects in amnestic and weakly trained rats. Physiological Psychology, 3, 265–269. DOI: 10.3758/BF03337522

Dickinson A, & Burke J (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. Quarterly Journal of Experimental Psychology, 36A, 29–50. DOI: 10.1080/713932614

Dunsmoor JE, Ahs F, Zielinski DJ, LaBar KS (2014). Extinction in multiple virtual reality contexts diminishes fear reinstatement in humans. Neurobiology of Learning and Memory. In press. DOI: 10.1016/j.nlm.2014.02.010

Ebbinghaus H (1964) Memory: A contribution to experimental psychology. (Ruger HA & Bussemus CE, Translators). New York. Dover (Original work published 1885). DOI: 10.5214/ans.0972.7531.200408

Estes WK (1955). Statistical theory of distributional phenomena in learning. Psychological Review, 62, 369–377. DOI: 10.1037/h0046888 [PubMed: 13254976]

Fast CD, & Blaisdell AP (2011). Rats are sensitive to ambiguity. Psychonomic Bulletin & Review, 18, 1230–1237. DOI: 10.3758/s13423-011-0171-0 [PubMed: 21968926]

Fazio LK, Argawal PK, Marsh EJ, & Roediger HL (2010). Memorial consequences of multiple-choice testing on immediate and delayed tests. Memory & Cognition, 38, 407–418. DOI: 10.3758/MC.38.4.407 [PubMed: 20516221]

Fazio LK, Huelser BJ, Johnson A, & Marsh EJ (2010). Receiving right/wrong feedback: Consequences for learning. Memory, 18, 335–350. DOI: 10.1080/09658211003652491 [PubMed: 20408043]

Finley JR, Benjamin AS, Hays MJ, Bjork RA, & Kornell N (2011). Benefits of accumulating versus diminishing cues in recall. Journal of Memory and Language, 64, 289–298. DOI: 10.1016/j.jml.2011.01.006 [PubMed: 21499516]

Finn B, & Roediger HL (2013). Interfering effects of retrieval in learning new information. Journal of Experimental Psychology: Learning, Memory, and Cognition, 39, 1665–1682. DOI: 10.1037/a0032377 [PubMed: 23565780]

Gates AI (1917). Recitation as a factor in memorizing. Archives of Psychology, 6(40).

Glautier S & Elgueta T (2009). Multiple cue extinction effect on recovery of responding in causal judgments. International Journal of Comparative Psychology, 22, 254–270. DOI: 67g5g8rf

Glover JA (1989). The "testing" phenomenon: Not gone but nearly forgotten. Journal of Educational Psychology, 81, 392–399. DOI: 10.1037/0022-0663.81.3.392

Glover JA, Krug D, Hannon S, & Shine A (2010). The "*testing*" effect and restricted retrieval rehearsal. Psychological Record, 40, 215–226, DOI: 10.1007/BF03399560

Goode MK, Geraci L, & Roediger HL III. (2008). Superiority of variable to repeated practice in transfer on anagram solution. Psychonomic Bulletin & Review, 15, 662–666. DOI: 10.3758/PBR.15.3.662 [PubMed: 18567271]

Gordon WC, & Mowrer RR (1980). An extinction trial as a reminder treatment following electroconvulsive shock. Animal Learning & Behavior, 8, 363–367. DOI: 10.3758/BF03199618

Goswick AE, Fazio LK, & Marsh EJ (2010). The effects of multiple-choice testing and feedback on school-aged children. Annual meeting of the Psychonomic Society. St. Louis, MO.

Gunther LM, Denniston JC, Miller RR (1998). Conducting exposure treatment in multiple contexts can prevent relapse. Behavior Research and Therapy, 36, 75–91. DOI: 10.1016/S0005-7967(97)10019-5

Gunther LM, Miller RR, & Matute H (1997). CSs and USs: What's the difference? Journal of Experimental Psychology: Animal Behavior Processes 23, 15–30. DOI: 10.1037/0097-7403.23.1.15 [PubMed: 9008860]

Halamish V, & Bjork RA (2011). When does testing enhance retrieval? A distribution-based interpretation of retrieval as a memory modifier. Journal of Experimental Psychology: Learning, Memory and Cognition, 37, 801–812. DOI: 10.1037/a0023219 [PubMed: 21480751]

Hattie J, & Timperley H (2007). The power of feedback. Review of Educational Research, 77, 81–112. DOI: 10.3102/003465430298487

Hogan RM, & Kintsch W (1971). Differential effects of study and test trials on long-term recognition and recall. Journal of Verbal Learning and Verbal Behavior, 10, 562–567. DOI: 10.1016/S0022-5371(71)80029-4

Izawa C (1967). Function of test trials in paired-associate learning. Journal of Experimental Psychology, 75, 194–209. DOI: 10.1037/h0024971 [PubMed: 6062960]

Izawa C (1970). Optimal potentiating effects and forgetting -prevention effects of tests in paired-associate learning. Journal of Experimental Psychology, 83, 340–344. DOI: 10.1037/h0028541

Jönsson FU, Kubik V, Sundqvist ML, Todorov I, & Jonsson B (2014). How crucial is the response format for the testing effect? Psychological Research, 78(5), 623–633. DOI: 10.1007/s00426-013-0522-8. [PubMed: 24173813]

Kamin LJ (1969). Predictability, surprise, attention and conditioning. In Campbell BA & Church RM (eds.), Punishment and aversive behavior, 279–96, New York: Appleton-Century-Crofts.

Kane JH, & Anderson RC (1978). Depth of processing and interference effects in the learning and remembering of sentences. Journal of Educational Psychology, 70, 626–635. DOI: 10.1037/0022-0663.70.4.626

Kang SHK, McDermott KB, & Roediger HL (2007). Test format and corrective feedback modulate the effect of testing on memory retention. European Journal of Cognitive Psychology, 19, 528–558. DOI: 10.1080/09541440601056620

Kantner J, & Lindsay SD (2010). Can corrective feedback improve recognition memory? Memory & Cognition, 38, 389–406. DOI: 10.3758/MC.38.4.389 [PubMed: 20516220]

Karpicke JD, Butler A, & Roediger HL (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? Memory, 17, 471–479. DOI: 10.1080/09658210802647009 [PubMed: 19358016]

Karpicke JD, Lehman M, & Aue WR (2014). Retrieval-based learning: An episodic context account. The psychology of learning and motivation: Vol. 61 (p. 237–284). Cambridge, MA: Elsevier Academic Press. DOI: 10.1016/B978-0-12-800283-4.00007-1

Karpicke JD, & Roediger HL (2007). Repeated retrieval during learning is the key to long-term retention. Journal of Memory and Language, 57, 151–162. DOI: 10.1016/j.jml.2006.09.004

Karpicke JD, & Roediger HL (2008). The critical importance of retrieval for learning. Science, 319, 966–968. DOI: 10.1126/science.1152408 [PubMed: 18276894]

Karpicke JD, & Zaromb FM (2010). Retrieval mode distinguishes the testing effect from the generation effect. Journal of Memory & Language, 62, 227–239. DOI: 10.1016/j.jml.2009.11.010

Knight JB, Ball BH, Brewer GA, DeWitt MR, Marsh RL (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. Journal of Memory and Language, 66, 731–746. DOI: 10.1016/j.jml.2011.12.008

Kolers PA (1973). Remembering operations. Memory & Cognition, 1, 347–355. DOI: 10.3758/BF03198119 [PubMed: 24214568]

Konorski J (1967). Integrative activity of the brain. Chicago, IL: University of Chicago Press. DOI: 1967–35012-000

Kornell N (2014). Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. Journal of Experimental Psychology: Learning, Memory, and Cognition, 40, 106–114. DOI: 10.1037/a0033699 [PubMed: 23855547]

Kornell N and Bjork RA (2007). The promise and perils of self-regulated study. Psychonomic Bulletin & Review, 14, pp. 219–224. DOI: 10.3758/BF03194055 [PubMed: 17694904]

Kornell N, Hays MJ, & Bjork RA (2009). Unsuccessful retrieval attempts enhance subsequent learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 35, 989–998. DOI: 10.1016/j.jml.2011.04.002 [PubMed: 19586265]

Kornell N, Bjork RA, & Garcia MA (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. Journal of Memory and Language, 65, 85–97. DOI: 10.1016/j.jml.2011.04.002

Laborda MA, & Miller RR (2013) Preventing return of fear in an animal model of anxiety: Additive effects of massive extinction and extinction in multiple contexts. Behavior Therapy, 44, 249–261. DOI: 10.1016/j.beth.2012.11.001 [PubMed: 23611075]

Landauer TK, & Bjork RA (1978). Optimum rehearsal patterns and name learning. In Gruneberg MM, Morris PE, & Sykes RN (Eds.), Practical aspects of memory (pp. 625–632). London: Academic Press. NAID: 10009702927

Logan JM, Thompson AJ, Marshak DW, (2011). Testing to enhance retention in human anatomy. Anatomical Sciences Education, 4, 243–248. DOI: 10.1002/ase.250 [PubMed: 21805688]

Marsh EJ, Agarwal PK, & Roediger HL III. (2009). Memorial consequences of answering SAT II questions. Journal of Experimental Psychology: Applied, 15, 1–11. DOI: 10.1037/a0014721 [PubMed: 19309212]

McDaniel MA, Agarwal PK, Huelser BJ, McDermott KB, & Roediger HL (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. Journal of Educational Psychology, 103, 399–414. DOI: 10.1037/a0021782

McDaniel MA, & Fisher RP (1991). Tests and test feedback as learning sources. Contemporary Educational Psychology, 16, 192–201. DOI: 10.1037/0278-7393.11.2.371

McDaniel MA, & Masson MEJ (1985). Altering memory representations through retrieval. Journal of Experimental Psychology: Learning, Memory, and Cognition, 11, 371–385. doi.org/10.1037/0278-7393.11.2.371

McDaniel MA, Kowitz MD, & Dunay PK (1989). Altering memory through recall: The effects of cue-guided retrieval processing. Memory & Cognition, 17, 423–434. DOI: 10.3758/BF03202614 [PubMed: 2761400]

McDaniel MA, Wildman KM, & Anderson JL (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. Journal of Applied Research in Memory and Cognition, 1, 18–26. DOI: 10.1016/j.jarmac.2011.10.001

McDermott KB, & Naaz F (2014). Is recitation an effective tool for adult learners? Journal of Applied Research in Memory and Cognition, 3(3), 207–213. DOI: 10.1016/j.jarmac.2014.06.006

McLaren IP, & Mackintosh NJ (2002). Associative learning and elemental representation: II. Generalization and discrimination. Animal Learning & Behavior, 30, 177–200. DOI: 10.3758/BF03192828 [PubMed: 12391785]

Miguez G, Laborda MA, Miller RR (2014). Enhancement and reduction of associative retroactive cue interference by training in multiple contexts. Learning & Behavior, 42, 318–329. DOI: 10.3758/s13420-014-0149-7 [PubMed: 25035103]

Miller RR (1982). Effects of intertrial reinstatement of training stimuli on complex maze learning in rats: Evidence that “acquisition” curves reflect more than acquisition. Journal of Experimental Psychology: Animal Behavior Processes, 8, 86–109. DOI: 10.1037/0097-7403.8.1.86

Miller RR, Ott CA, Berk AM, & Springer AD (1974). Appetitive memory restoration after electroconvulsive shock in the rat. Journal of Comparative Physiological Psychology, 87, 717–723. DOI: 10.1037/h0036972 [PubMed: 4426993]

Miller RR, & Matzel LD (1988). The comparator hypothesis: A response rule for the expression of associations. In Bower GH (Ed.), The psychology of learning and motivation (Advances in research and theory, Vol. 22, pp. 51–92). San Diego: Academic Press. DOI: 10.1016/S0079-7421(08)60038-9

Monfils MH, Cowansage KK, Klann E, & LeDoux JE (2009). Extinction-reconsolidation boundaries: Key to persistent attenuation of fear memories. Science, 324, 951–955. DOI: 10.1126/science.1167975 [PubMed: 19342552]

Morris CD, Bransford JD, & Franks JJ (1977). Levels of processing versus transfer-appropriate processing. Journal of Verbal Learning and Verbal Behavior, 16, 519–533. DOI: 10.1016/S0022-5371(77)80016-9

Mowrer RR, & Gordon WC (1983). The effects of cuing in an “irrelevant” context. Animal Learning & Behavior, 11, 401–406. DOI: 10.3758/BF03199794

Nader K, Schafe GE, & Le Doux JE (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. Nature, 406(6797), 722–726. DOI: 10.1038/35021052 [PubMed: 10963596]

Pan SC, & Sana F (2021). Pretesting versus posttesting: Comparing the pedagogical benefits of errorful generation and retrieval practice. Journal of Experimental Psychology: Applied, 27(2), 237–257. 10.1037/xap0000345 [PubMed: 33793291]

Pashler H, Cepeda NJ, Wixted JT, & Rohrer D (2005). When does feedback facilitate learning of words? Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 3–8. DOI: 10.1037%2F0278-7393.31.1.3 [PubMed: 15641900]

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. DOI: 10.1037/0278-7393.29.6.1051 [PubMed: 14622045]

Pavlov IP (1927). Conditioned reflexes. (Anrep GV, Ed. & Trans.) London: Oxford University Press. DOI: 10.5214/ans.0972-7531.1017309

Pearce JM (1987). A model for stimulus generalization in Pavlovian conditioning. Psychological Review, 94, 61–73. DOI: 10.1037/0033-295X.94.1.61 [PubMed: 3823305]

Pearce JM, & Hall G (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychological Review, 87, 532–552. DOI: 10.1037/0033-295X.87.6.532 [PubMed: 7443916]

Peterson DJ, & Mulligan NW (2013). The negative testing effect and multifactor account. Journal of Experimental Psychology: Learning, Memory, and Cognition, 39, 1287–1293. DOI: 10.1037/a0031337 [PubMed: 23421505]

Pierce BH, Gallo DA, & McCain JL (2017). Reduced interference from memory testing: A postretrieval monitoring account. Journal of Experimental Psychology: Learning, Memory, and Cognition, 43(7), 1063–1072. DOI: 10.1037/xlm0000377 [PubMed: 28114777]

Potts R, & Shanks DR (2012). Can testing immunize memories against interference? Journal of Experimental Psychology: Learning, Memory, and Cognition, 38(6), 1780–1785. DOI: 10.1037/a0028218 [PubMed: 22686838]

Potts R, & Shanks DR (2014). The benefit of generating errors during learning. Journal of Experimental Psychology: General, 143(2), 644. DOI: 10.1037/a0033194 [PubMed: 23815457]

Pyc MA, & Rawson KA (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. DOI: 10.1016/j.jml.2009.01.004

Racsmany M, Conway MA, & Demeter G (2010) Consolidation of episodic memories during sleep: Long-term effects of retrieval practice. Psychological Science, 21, 80–85. DOI: 10.1177/0956797609354074 [PubMed: 20424027]

Rescorla RA, & Heth CD (1975). Reinstatement of fear to an extinguished conditioned stimulus. Journal of Experimental Psychology: Animal Behavior Processes, 1, 88–96. DOI: 10.1037/0097-7403.1.1.88 [PubMed: 1151290]

Rescorla RA, & Wagner AR (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In Black AH & Prokasy WF (Eds.), Classical conditioning II: Current research and theory (pp. 64–99). New York: Appleton-Century-Crofts. NAID: 10024275851

Roediger HL (1990). Implicit memory: Retention without remembering. American Psychologist, 45, 1043–1056. DOI: 10.1037/0003-066X.45.9.1043 [PubMed: 2221571]

Roediger HL III, (2008). Relativity of remembering: Why the laws of memory vanished. Annual Review of Psychology. 59, 225–254. DOI: 10.1146/annurev.psych.57.102904.190139

Roediger HL, Agarwal PK, Kang SHK, & Marsh EJ (2009). Benefits of testing memory: Best practices and boundary conditions. In Davies GM & Wright DB (Eds.), Current issues in applied memory research (pp. 13–49). Hove, U.K.: Psychology Press. DOI: 10.4324/9780203869611-10

Roediger HL, & Karpicke JD (2006a). Test enhanced learning: Taking memory tests improves long-term retention. Psychological Science, 17, 249–255. DOI: 10.1111/j.1467-9280.2006.01693.x [PubMed: 16507066]

Roediger HL, & Karpicke JD (2006b). The power of testing memory: Basic research and implications for educational practice. Perspectives on Psychological Science, 1, 181–210. DOI: 10.1111/j.1745-6916.2006.00012.x [PubMed: 26151629]

Roediger HL, Putnam AL, & Smith MA (2011). The benefits of testing and their applications to educational practice. In Mestre J & Ross BH (Eds.), The psychology of learning and motivation: Cognition in education (pp.1–36). Amsterdam, The Netherlands: Elsevier. DOI: 10.1016/B978-0-12-387691-1.00001-6

Roediger HL, & Smith MA (2012). The “pure-study” learning curve: The learning curve without cumulative testing. Memory & Cognition, 40, 989–1002. DOI: 10.3758/s13421-012-0213-5 [PubMed: 22644774]

Rohrer D, Taylor K, & Sholar B (2010). Tests enhance the transfer of learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36, 233–239. DOI: 10.1037/a0017678 [PubMed: 20053059]

Rosenbaum DA, Carlson RA, & Gilmore RO (2001). Acquisition of intellectual and perceptual-motor skills. Annual Review of Psychology, 52, 453–470. DOI: 10.1146/annurev.psych.52.1.453

Rowland CA, & DeLosh EL (2014). Mnemonic benefits of retrieval practice at short retention intervals. Memory, 23, 403–419. doi.org/10.1080/09658211.2014.889710 [PubMed: 24579674]

Schmidt RA, Young DE, Sinnen S, & Shapiro DC (1989). Summary knowledge of results for skill acquisition: Support for the guidance hypothesis. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 352–359. DOI: 10.1037/0278-7393.15.2.352 [PubMed: 2522520]

Skaggs EB (1920). The relative value of grouped and interspersed recitations. Journal of Experimental Psychology, 3, 424–446. DOI: 10.1037/h0072486

Skinner BF (1958). Teaching machines. Science, 128, 969–977. DOI: 10.1126/science.128.3330.969 [PubMed: 13592277]

Smith MA, Roediger HL, & Karpicke JD (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. Journal of Experimental Psychology: Learning, Memory, and Cognition, 39, 1712–1725. DOI: 10.1037/a0033569 [PubMed: 23815513]

Smith TA, & Kimball DR (2010). Learning from feedback: spacing and the delay-retention effect. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36, 80–95. DOI: 10.1037/a0017407 [PubMed: 20053046]

Spitzer HF (1939). Studies in retention. Journal of Educational Psychology, 30, 641–656. DOI: 10.1037/h0063404

Surprenant AM, & Neath I (2009). Principles of memory. New York: Psychology Press.

Storm BC, Friedman MC, Murayama K, Bjork RA (2014) On the transfer of prior tests or study events to subsequent study. Journal of Experimental Psychology: Learning, Memory, and Cognition, 40, 115–124. DOI: 10.1037/a0034252 [PubMed: 23978234]

Stout SC, & Miller RR (2007). Sometimes competing retrieval (SOCR): A formalization of the extended comparator hypothesis. Psychological Review, 114, 759–783. DOI: 10.1037/0033-295X.114.3.759 [PubMed: 17638505]

Thompson CP, Wenger SK, & Bartling CA (1978). How recall facilitates subsequent recall: A reappraisal. Journal of Experimental Psychology: Human Learning and Memory, 4, 210–221. DOI: 10.1037/0278-7393.4.3.210

Thorndike EL (1898). Animal intelligence: an experimental study of the associative processes in animals. The Psychological Review: Monograph Supplements, 2, i–109. DOI: 10.1037/h0092987

Tolman EC (1932). Purposive behavior in animals and men. New York: Century.

Tullis JG, Fiechter JL, Benjamin AS (2017). The efficacy of learners' testing choices. Journal of Experimental Psychology: Learning, Memory and Cognition 44(4), 540–552. DOI: 10.1037/xlm0000473 [PubMed: 29094989]

Tulving E (1967). The effects of presentation and recall of material in free-recall learning. Journal of Verbal Learning and Verbal Behavior, 6, 175–184. DOI: 10.1016/S0022-5371(67)80092-6

Van Hamme LJ, & Wasserman EA (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. Learning and Motivation, 25, 127–151. DOI: 10.1006/lmot.1994.1008

Vansteenwegen D, Vervliet B, Iberico C, Baeyens F, Van den Bergh O, & Hermans D (2007). The repeated confrontation with videotapes of spiders in multiple contexts attenuates renewal of fear in spider anxious students. Behavior Research and Therapy, 45,1169–1179. DOI: 10.1016/j.brat.2006.08.023

Wagner AR (1981). SOP: A model of automatic memory processing in animal behavior. In Spear NE & Miller RR (Eds.), Information processing in animals: Memory mechanisms (pp. 5–47). Hillsdale, NJ: Erlbaum. DOI: 10.4324/9781315798820

Waldmann MR, Schmid M, Wong J, & Blaisdell AP (2012). Rats distinguish between absence of events and lack of evidence in contingency learning. Animal Cognition, 15, 979–990. DOI: 10.1007/s10071-012-0524-8 [PubMed: 22744612]

Wasserman EA, & Castro L (2005). Surprise and change: Variations in the strength of present and absent cues in causal learning. Learning & Behavior, 33, 131–146. DOI: 10.3758/BF03196058 [PubMed: 16075834]

Wheeler DS, & Miller RR (2007). Primacy effects induced by temporal or physical context shifts are attenuated by a preshift test trial. Quarterly Journal of Experimental Psychology, 60,191–210. DOI: 10.1080/17470210600790240

Wheeler MA, Ewers M, & Buonanno J (2003). Different rates of forgetting following study versus test trials. Memory, 11, 571–580. DOI: 10.1080/09658210244000414 [PubMed: 14982124]

Yang C, Luo L, Vadillo MA, Yu R, & Shanks DR (2021, March 8). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. Psychological Bulletin, Advance online publication. doi.org/10.1037/bul0000309