The "pure-study" learning curve: The learning curve without cumulative testing

Henry L. Roediger III · Megan A. Smith

Published online: 30 May 2012

Abstract The customary assumption in the study of human learning using alternating study and test trials is that learning occurs during study trials and that test trials are useful only to measure learning. In fact, tests seem to play little role in the development of learning, because the learning curve is similar even when the number of test trials varies widely (Tulving, Journal of Verbal Learning and Verbal Behavior 6:175–184, 1967). However, this outcome seems odd, because other research has shown that testing fosters greater long-term learning than does studying. We report three experiments addressing whether tests affect the shape of the learning curve. In two of the experiments, we examined this issue by varying the number of spaced study trials in a sequence and examining performance on only a single test trial at the end of the series (a "pure-study" learning curve). We compared these pure-study learning curves to standard learning curves and found that the standard curves increase more rapidly and reach a higher level in both free recall (Exp. 1) and paired-associate learning (Exp. 2). In Experiment 3, we provided additional study trials in the "pure-study" condition to determine whether the standard (study–test) condition would prove superior to a study–study condition. The standard condition still produced better retention on both immediate and delayed tests. Our experiments show that test trials play an important role in the development of learning using both free-recall (Exps. 1 and 3) and paired-associate (Exp. 2) procedures. Theories of learning have emphasized processes that occur during study, but our results show that processes engaged during tests are also critical.

H. L. Roediger III (✉) · M. A. Smith
Department of Psychology, Box 1125, Washington University,
One Brookings Drive,
St. Louis, MO 63130-4899, USA
e-mail: roediger@wustl.edu

Keywords Learning · Memory · Testing effect · Recall

During the 20th century, learning was perhaps the central focus of experimental psychology, with experiments performed on rats, mice, cats, pigeons, dogs, monkeys, and humans (among other creatures). In the study of human learning, researchers beginning with Ebbinghaus (1885/1964) used various arrangements of multiple study–test procedures with lists of nonwords or words. One common procedure used to study learning within this tradition—the study–test method—is the focus of this article. In the study–test procedure, subjects first study a set of material and are tested on it, then they study it again (either in the same order or in a new, random order) and take a second test, and so on, for as many trials as desired (or, sometimes, until the subjects reach a specified criterion). The resulting function relating the number of learning trials (on the abscissa) to performance on some dependent measure (on the ordinate) is the *learning curve*. For most tasks, the learning curve is negatively accelerated, although debate exists as to the function that best fits and whether the various functions fit because of averaging artifacts (e.g., Heathcote, Brown, & Mewhort, 2000; Mazur & Hastie, 1978). Exponential and power functions are the primary contenders for such curves, but the essential point for the present purposes is that in both functions, learning develops rapidly over early trials and then slows markedly, even when subjects are not near ceiling-level performance. Learning curves (like forgetting curves) show an impressive similarity across many (but not all) tasks. Some authoritative reviews on human learning have been provided by McGeoch (1942), Hovland (1951), and Estes (1988), among others.

Surprisingly, the study of human learning has waned in the field known as "human learning and memory," which today consists mostly of the study of memory. That is, most experiments today are single-trial affairs, with a single study and test phase and the emphasis on memory for the study material, as assessed on the single test. At the risk of seeming retrograde, our study reopens a puzzle in the study of learning and seeks to solve it.

Many theories of human learning have been proposed over the years. Ebbinghaus (1885/1964) assumed that repetitions of information create memory traces, that these traces of experience vary in strength, and that strength cumulates gradually over repeated presentations. Others have generally followed his lead and have also assumed that learning causes traces to grow gradually in strength (e.g., Hull, 1952; Underwood & Keppel, 1962). The implicit assumption in theories of learning using the study–test method has been that learning occurred on study trials and its effects were simply displayed on test trials. The usual assumption (again implicitly held, from the omission of any discussion of tests playing a role) was that the study of events increased their trace strength (Underwood & Kepple, 1962), or sometimes the number of traces (Bernbach, 1970). An even more radical view was promulgated some 50 years ago by Rock (1957; Rock & Heimer, 1959) and by Estes (1960; Estes, Hopkins, & Crothers, 1960). They suggested that even during study trials, only a few items transition from an unlearned state to a learned state, and that the smooth learning curve is essentially an averaging artifact. They argued that learning in multitrial experiments occurred in an all-or-none fashion; that is, the trace of an event either gained 100% of its strength on a single trial, or none at all. In recently reviewing this debate, Roediger and Arnold (2012) found no mention at all of test trials affecting learning. The emphasis was on whether learning grew gradually or in an all-or-none manner during study trials in paired-associate learning. Test trials seem to have been assumed to simply measure the learning occurring during the study trials. At the very least, no theories mentioned a role for testing during the 1950s and early 1960s (but see Miller & McGill, 1952, for a possible exception). Of course, this general assumption agrees with the one embedded in our educational system: Learning occurs from various study activities (reading, lecturing, reviewing, outlining, etc.), and tests (which are often given rather infrequently) measure the learning that has occurred from these study activities.

Investigators during the past 20 years have systematically explored the effects of taking tests on retention, noting both positive effects (the *testing effect*; Carrier & Pashler, 1992) and negative effects (*retrieval-induced forgetting*; Anderson, Bjork, & Bjork, 1994)—although both of these endeavors have an earlier history (e.g., Gates, 1917; Roediger, 1974). The present article focuses on the positive effects of testing, or the testing effect (sometimes called the *retrieval practice effect*). The basic finding is that subjects who have been given a test and successfully retrieved information remember this information better on a later test, relative to either of two comparison conditions: subjects taking no initial test or subjects restudying the same material in place of an initial test (e.g., Carrier & Pashler, 1992; Halamish & Bjork, 2011; Pyc & Rawson, 2009; Wheeler & Roediger, 1992, among many others). In general, taking a test covering recently studied material has large positive effects relative to either control condition. This is especially true if the tests involve feedback (for when retrieval failures occur on the first test) and when the tests are delayed by a day or more. In some cases, the effects can be surprisingly large (Karpicke & Roediger, 2008). A thorough historical review of the testing effect can be found in Roediger and Karpicke (2006a), and more recent ones are in Roediger and Butler (2011) and Roediger, Putnam, and Smith (2011).

The aim of this study is to solve a mystery involving the testing effect and its relation to the development of learning over trials. Briefly, over the past 45 years, many multitrial free-recall learning experiments have seemed to show that the amount of testing involved in a learning experiment does not increase learning. Because testing has such powerful effects in many situations, the absence of any positive effects in the development of learning seems odd. We first review the background for the mystery (and show that even the existence of this "mystery" represents a kind of hindsight bias on our part), and then we present three experiments to solve it.

Izawa (1967) and Tulving (1967) were the first to ask about the relative influences of study and test trials in learning experiments. Izawa (1967, 1971) emphasized the potentiating effects of taking a test on new study opportunities. She showed that subjects learned more from a second study opportunity if they had been tested after a first study trial than if they had not been tested, and she referred to this benefit as *test-potentiated learning*. In a multitrial procedure, subjects should show greater potentiation on future study trials from receiving tests. Tulving (1967) explored the relative effects of study and test trials directly by exploring the role of other sequences of study and test trials, besides the usual alternating study and test phases (denoted *STST*). Following Tulving (1967), we will denote a sequence of four study and test trials as a cycle, so STST represents one cycle in a standard learning experiment, and we consider it the standard or baseline condition for purposes of comparison with two other types of cycles. If learning occurs only on study trials, and tests merely measure the preceding learning, then changes to the standard learning cycle should have profound effects. Tulving (1967) developed two new types of study–test cycles: SSST and STTT. If learning occurs during study trials and not test trials, then relative to the standard cycle, the SSST cycle should enhance recall (there is an extra study trial inserted in place of a test trial). In contrast, the STTT cycle should harm recall (due to only one study trial being in each cycle, rather than two).

Tulving (1967) had subjects learn a list of 36 words under conditions of free recall using each of these three conditions (STST, SSST, and STTT), with 36 s for study and 36 s for tests (with oral recall). Thus, the study and test periods were equated in time. Of course, during the study periods, all 36 items were presented, whereas during the test periods, only the items that could be recalled were reexperienced; thus, this factor would seem to strongly favor the study conditions over the test conditions in learning (and, hence, also to favor conditions with more study trials). Six cycles of four trials were given in each of the three conditions (STST, SSST, and STTT), so that overall the numbers of study trials were 12, 18, and 6, respectively, in the three conditions. Thus, under the assumption that learning occurs during study trials, one should predict that during the course of learning (and certainly by the end of learning), the ordering of conditions in terms of performance would be SSST > STST > STTT. However, Tulving (1967) showed that learning curves from the three conditions looked surprisingly similar throughout the 24 trials. Even on the 24th trial, which was always a test trial, no differences appeared among the conditions, despite the fact that some subjects had studied the list 18 times, and others only six. Again, this lack of difference occurred despite the fact that the study trials offered 100% reexposure to the list, whereas the test trials did not. Tulving (1967) concluded that recall after a study phase "depends primarily on the total amount of time spent on the task, and that it is relatively little affected, if at all, by the distribution of this time between studying and recalling the material. This finding clearly implies that a recall test in FRL [free-recall learning] serves other functions beside that of measuring the amount or degree of learning" (1967, p. 181).

Tulving's (1967) results were remarkable in showing that tests were just as powerful in influencing learning as were study trials. Of course, bearing Izawa's (1971) hypothesis of test potentiation in mind, it could be argued that the large number of test trials in the STTT sequences caused subjects to gain more from their relatively few study trials, and that the test-potentiating effect balanced the greater number of study trials in other conditions. This occurrence seems unlikely, but there is no way to know. In the late 1960s and early 1970s, several sets of investigators replicated and extended Tulving's (1967) work, and they generally confirmed his conclusion (Bregman & Wiener, 1970; Donaldson, 1971; Lachman & Laughery, 1968; Patterson, 1972; Rosner, 1970). The issue lay fallow for many years, but Karpicke and Roediger (2007) reopened it with two experiments that mostly replicated Tulving's (1967) conclusions concerning the effects of study and test trials during learning, except that the alternating study–test trials led to somewhat greater learning in their procedure (which differed in some details from that of Tulving, 1967). Still, the condition with the largest number of study trials—SSST—did not lead to the greatest learning, contrary to the idea that study trials lead to superior learning relative to test trials. Karpicke and Roediger (2007) showed that when retention tests were given a week later, the STTT condition led to greater recall than did the SSST condition, again showing the power of test trials relative to study trials in enhancing long-term retention (see also Karpicke & Roediger, 2008).

The emphasis in Tulving's (1967) research and the studies that succeeded it focused on the remarkable fact that a test trial (even one on which recall is far from perfect) can have as much impact on learning as a study trial (with 100% reexposure of items). Of course, Tulving's (1967) experiment occurred long before the current burst of activity on testing. With the pure 20/20 wisdom of hindsight, we can now turn Tulving's (1967) question on its head and ask, if test trials are so important for long-term retention, why don't they produce *more* learning than study trials, even during the learning phase of the experiment? Tulving (1967) raised this issue himself, briefly, citing the testing (or recitation) effects shown by Gates (1917). Izawa's (1971) studies showing test-potentiated learning led to the same puzzle. Two methodological answers to this question seem possible, and depending on which is correct, different theoretical implications also follow.

First, a number of studies have shown differences in the effects of study and test trials in immediate and delayed retention. For example, Hogan and Kintsch (1971) had subjects study a list of words four times (SSSS) or study it once and take three tests (STTT)—one cycle for each condition, in Tulving's (1967) terms. On an immediate test (in the same session), subjects in the repeated-study condition recalled more words. However, on a delayed test the effect reversed, and the STTT condition showed greater recall than the SSSS condition. This same pattern of results has been shown by Roediger and Karpicke (2006b), Thompson, Wenger, and Bartling (1978), and Wheeler, Ewers, and Buonanno (2003). Thus, one possibility is that (for whatever reason) it takes time for the power of test trials to be revealed, and that on tests in the same session, study events produce better recall than do test events. In terms of the Bjorks' "new theory of disuse" (Bjork & Bjork, 1992), which distinguishes between retrieval strength (i.e., factors affecting the temporary accessibility of information over the short term) and storage strength (i.e., factors affecting long-term storage), the effect of testing is greater on storage strength and less on retrieval strength, whereas study trials have the opposite effect (Halamish & Bjork, 2011). Thus, within this theory, it may well be that repeated study may boost temporary retrieval strength, whereas testing enhances long-term storage strength.

On the other hand, many studies showing positive effects of testing have been conducted in a single session (e.g., Carrier & Pashler, 1992, among many others), so the first option outlined may not be correct. In addition, if study trials always produced better performance on immediate tests, in Tulving's (1967) and Karpicke and Roediger's (2007) experiments, the SSST condition should have greatly outpaced the STTT condition across learning trials, because both conditions involved relatively immediate tests. This outcome did not occur, and performance on those two conditions was roughly comparable.

These facts lead to a second possible interpretation of why test trials do not appear to lead to greater learning than do study trials in the standard learning experiment. The hypothesis is that test trials actually do lead to greater learning than study trials in the learning experiments, but that only a few test trials are needed to enhance acquisition. That is, in all prior experiments, beginning with Tulving (1967), all experimental conditions have involved a mix of study and test trials. For example, as noted, in the SSST condition of Tulving's (1967) original experiment, six test trials occurred across the six cycles, even in the condition with the largest number of study trials. Perhaps these six trials permitted enough benefit from testing that, with added study trials, performance during learning was as good as in the other conditions (STST and STTT), with 12 and 18 test trials, respectively.

These considerations led us to ask: What would initial learning look like if it were possible to completely remove test trials and examine the impact of only study trials on the learning curve? Would a "pure-study" learning curve develop in the same way as the standard curve arising from the usual study–test procedure? No research, to our knowledge, has completely removed all test trials, and with good reason—test trials are necessary to assess the learner's retention following study trials. But we can entertain the question, what would a learning curve based only on study trials look like relative to a standard learning curve? Theories of learning, as well as prior research (Karpicke & Roediger, 2007; Tulving, 1967), lead to the prediction of no difference between a "pure-study" learning curve and the standard function developed from repeated study–test trials. After all, at least in the tradition of human learning research, learning curves have been discussed for over a century as if they arose only from processes occurring during the study phases of learning experiments.

We designed a procedure to answer these questions by taking a standard STST learning procedure and removing all except the last (criterial) test trial after a varied number of study trials, to measure learning from only previous study episodes. We accomplished this "pure-study" learning curve by removing tests from the standard study–test sequence of a learning trial and replacing them with filled intervals of equal duration (see Table 1). The intervals were filled with a nonverbal task (Pac-Man) that we assumed would be dissimilar from the lists to be learned and would not cause interference (e.g., Roediger, Knight, & Kantowitz, 1977; Roediger & Payne, 1982). We used a within-subjects design, such that all subjects learned six different lists that were varied in terms of the numbers of study (and test) events that occurred prior to a final test.

Table 1 illustrates the conditions. As noted, when we removed test phases, we replaced them with a filler task to hold the spacing between each study trial constant across conditions (because spaced repetitions may themselves be partly responsible for good performance in the usual study–test sequence of a learning trial; Melton, 1970; see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). Following the last study trial in the pure-study conditions, subjects completed one test trial to assess their level of recall. As shown in Table 1 for the Study 2–8 conditions, the number of study phases completed before taking the test differed for each list, so that we were able to measure subjects' performance after a varying number of study trials. Performance from the single tests in the Study 2–8 conditions was plotted in order to create the pure-study learning curve. We also included the standard multitrial learning procedure (with alternating study and test phases) to create the standard learning curve. If the Study 2–8 conditions swept out a pure-study learning curve that was equivalent to the standard learning curve, this outcome would indicate that the conclusions from the

Table 1 Experiment 1 design

	Period 1	Period 2	Period 3	Period 4	Period 5	Period 6	Period 7	Period 8
Standard	S T	S T	S T	S T	S T	S T	S T	S T
Study 8	S –	S –	S –	S –	S –	S –	S –	S T
Study 6	S –	S –	S –	S –	S –	S T		
Study 4	S –	S –	S –	S T				
Study 3	S –	S –	S T					
Study 2	S –	S T						

S denotes a study trial, T denotes a test trial, and – denotes a filler trial studies by Tulving (1967) and Karpicke and Roediger (2007), among others, are correct: Test trials and study trials are generally equivalent in producing learning. On the other hand, if the standard multitrial learning condition produced greater learning than the pure-study learning curve, the outcome would show that cumulative tests in the standard condition produced greater learning than an equal number of study trials without tests. The standard learning curve, of course, expresses the combined benefits of both study and test trials. If there were differences between the curves, we could conclude that including at least some test trials in the learning procedure is important to enhance learning.

In Experiment 1, we used the above logic in a free-recall procedure. In Experiment 2, the same general logic and procedure were used in a paired-associate learning paradigm. In addition, in Experiment 2 we measured retention on a final test after the initial learning procedure. In Experiment 3, we showed that the differences between the pure-study learning curve and the standard learning curve could not be explained merely by differential exposure to the studied material occasioned by the tests. We did this by filling the blank intervals in Table 1 (Study 2–8 conditions) with another presentation of the list. In all three experiments, we showed that the standard study–test learning condition was superior to the pure-study conditions.

Experiment 1

Method

Subjects A group of 48 subjects was recruited from the Washington University in St. Louis human subject pool and participated in exchange for partial course credit or payment ($10/h).

Materials The materials consisted of six lists composed of 40 concrete English nouns. All of the words were drawn from the MRC Psycholinguistic Database, and the characteristics for each word were obtained from the English Lexicon Project (Balota et al., 2007). All words were four to eight letters long, and the six lists were equated for word length, concreteness, and frequency.

Design Each subject completed six within-subjects learning conditions: standard and repeated study (two, three, four, six, and eight times). The design is illustrated in Table 1. In the standard learning condition, the subjects alternated between studying and recalling the list of words for a total of eight study trials and eight test trials. In the repeated-study learning conditions, subjects studied the list of words repeatedly for a given number of study trials, each separated by a filler task (playing Pac-Man). Immediately following the last study trial, the subjects completed a test trial. Twelve versions of the design were created, such that the ordering of conditions was counterbalanced and the subjects could not infer the pattern of the study and test trials that they were to complete for a given list. The order of list presentation was held constant across subjects, but each condition was paired with each list of words an equal number of times.

Procedure The subjects were tested individually or in small groups of four or fewer and were told that they would study and recall lists of words throughout the experiment. They were also told that they would study lists various numbers of times and would be tested on them every so often. The subjects learned one list for each of the six within-subjects conditions described above. Because of the length of the procedure (about 2.5 h), the experiment was separated into two sessions: The subjects learned three lists during the first session in three of the six conditions, and then learned the other three lists during the second session (which occurred from 3 to 14 days later). Each session lasted no more than 90 min.

The subjects proceeded through the series of study, test, and filler trials necessary for each condition. During study trials, we instructed the subjects to study the words so that they would be able to recall them later. Before each study trial, a “Begin Study” prompt appeared on the screen for 2 s so that subjects could prepare to study the list of words. Then 40 words were presented on the computer screen, one at a time at a 3-s rate. The words were presented in a random order (determined by the computer) on each trial.

During tests we instructed subjects to write down as many words as they could from the current list on the provided recall sheets. They were warned to recall only words from the list that they had most recently studied and not from prior lists. A “Begin Test” prompt appeared on the screen at the start of all test trials, indicating that subjects should begin recalling as many words as they could during the test period. In addition, the prompt indicated which recall sheet they should use (e.g., “Begin Test 1” indicated that subjects should recall words on the recall sheet labeled “Test 1”). All recall sheets were provided for the subjects before the experiment began in the order that they would need them. The test prompt remained on the screen for the duration of the 2-min test. When 2 min had passed, the computer screen flashed red, and the subjects were instructed to hand their recall sheets to the experimenter. We collected the recall sheets so that subjects could not look back to other tests at any time during the experiment. After the subjects turned in their recall sheets, they pressed Enter and proceeded to the next study trial. When the last test was complete for each condition, a “Begin New List” prompt appeared so that subjects were aware that it was time to learn a new list of words.

In the repeated-study conditions, subjects participated in the filler task between each study trial during the time when the standard group was tested, in order to equate the spacings of study in the two conditions. A "Begin Pac Man" prompt appeared for 2 s, and then subjects played Pac-Man for 2 min before moving on to the next study trial. Pac-Man was selected as a nonverbal task that we thought would prevent rehearsal but not cause interference in remembering the list. After subjects completed the six conditions, they were thanked and debriefed.

Results

All results in this and the other experiments were reliable at the .05 level of confidence, unless otherwise noted. A Greenhouse–Geisser correction was used for violations of the sphericity assumption in repeated measures analyses (Geisser & Greenhouse, 1958). On each test, each item was counted as correct if it varied from the list word only in singular/plural form or due to a slight misspelling, but no other variations of the items were accepted.

Figure 1 shows free-recall performance as a function of the number of study trials for both the standard and repeated-study conditions. In the standard condition, with alternating study and test phases, the function plots performance in the eight sequential test phases, as usual. The pure-study learning curve was created from each of the single test trials in the repeated-study conditions (after two, three, four, six, and eight study trials). Performance from the first test in the standard condition was used in both curves because they portrayed the same condition (one study trial prior to taking the free-recall test). As can be seen in Fig. 1, the standard learning condition with both study and test phases was clearly superior to the pure-study learning curve, in which only study phases occurred until the single test.

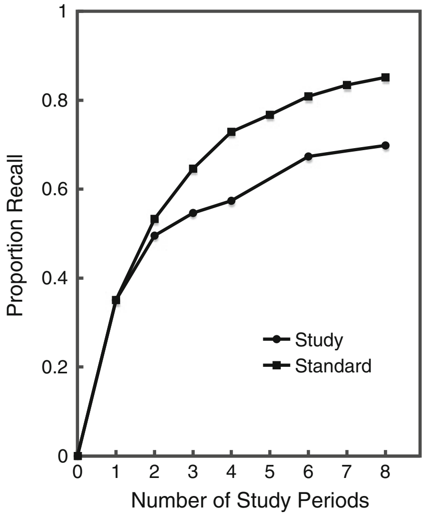

Fig. 1 Standard and pure-study free-recall learning curves in Experiment 1

We performed a 2 (condition: standard vs. repeated study) × 5 (trials: two, three, four, six, or eight) analysis of variance (ANOVA) adjusted for unequal trial intervals on the free-recall data. Only data from those trials in which a test was completed in both the standard and repeated-study conditions were included in the ANOVA (i.e., Tests 5 and 7 from the standard learning condition were not included). Of course, performance increased over trials in both the standard and repeated-study conditions [$F(4,\ 188) = 132.07$, $\eta_p^2 = .74$]. However, performance in the standard learning condition was greater than performance in the repeated-study learning condition [$F(1,\ 47) = 6.55$, $\eta_p^2 = .12$]; thus, practicing retrieval during the successive tests in the standard condition enhanced free-recall learning. Most importantly, we found an interaction between trial and condition [$F(4,\ 188) = 31.95$, $\eta_p^2 = .41$] indicating that across study trials, performance increased at a faster rate in the standard learning condition relative to the repeated-study learning conditions.

Discussion

The results show unequivocally that testing in the standard free-recall learning paradigm, with repeated study and test phases, does boost performance. Relative to the pure-study condition, the standard condition showed greater performance overall and greater learning over repeated study trials, as indicated by the condition-by-trial interaction. These findings seem to disagree with those of others (e.g., Karpicke & Roediger, 2007; Tulving, 1967), which showed that study trials and test trials are interchangeable in free-recall learning. However, in those experiments, all conditions included some test trials during the learning process, and not just a single final test, as in the repeated-study conditions in the present research. In addition, in those experiments, the number of study–test trials was held constant across conditions. Apparently, judging from prior work relative to the present results, only a few test trials are necessary to bring out learning in a free-recall paradigm.

Experiment 2

The results indicated that the standard free-recall learning curves depicted in the literature are expressing benefits of both study and test trials in concert. In Experiment 2, we asked how much learning is due to testing in paired-associate learning. Previous work has suggested that test trials play a greater role than do study trials in learning of paired associates (Karpicke, 2009, Exp. 3). However, when Bregman and Wiener (1970) compared studying and testing in both free recall and cued recall, they concluded that testing facilitates learning in free recall more than in cued recall. We wished to employ our repeated-study method to create a pure-study learning curve in paired-associate learning and to compare it to a learning curve derived in the standard way, with alternating study and test trials. Accordingly, Experiment 2 was conducted to replicate the results of Experiment 1, using the same general method with paired-associate learning. Five study trials (in the repeated-study condition) or five study–test trials (in the standard condition) were used rather than the eight trials used in Experiment 1.

Method

Subjects A group of 24 subjects was recruited from the Washington University in St. Louis human subject pool and participated in exchange for partial course credit or payment. None of the subjects had participated in Experiment 1.

Materials The materials consisted of six lists of English–English word pairs. Each list was composed of 100 words drawn from the English Lexicon Project (Balota et al., 2007), which were randomly paired to create 50 word pairs per list. Each word was four to eight letters long and had medium frequency (each list had an average hyperspace analogue to language [HAL] frequency ranging from 8,218.9 to 8,948.5; Lund & Burgess, 1996). Lists were equated for length and frequency across cue and target words. All of the materials were presented via computer.

Design The design was identical to that of Experiment 1, with two alterations. In the standard learning condition, the subjects alternated between study and test for a total of five study trials and five test trials. In the repeated-study learning condition, subjects studied for one, two, three, four, or five study trials prior to the single test trial. All learning conditions were manipulated within subjects by using six lists.

Procedure The procedure was similar to that of Experiment 1. The subjects proceeded through study trials, test trials, and the filler task in the same way as in Experiment 1. During a study trial, subjects were presented with pairs on the screen one at a time (e.g., “soccer–piano”) for 3 s, with the first word above the second word on the screen. During test trials, the subjects were presented with the top members of the pairs as cues with a cursor below them, and they were asked to type in the corresponding target words. Each cue was presented for 6 s, after which the computer advanced to the next cue, regardless of whether the subject had entered a response. Cues were randomly presented on each test trial in an order determined by the computer. After subjects had been tested on all 50 pairs, the computer advanced subjects to the next study trial in the standard learning condition. Between study trials during the repeated-study conditions, the subjects completed the Pac-Man filler task for 5 min. Subjects completed all conditions during one session, but after each list they were given an opportunity to take a break.

After the subjects completed all six of the initial learning conditions, they were once again given the opportunity to take a break. Then they took a final cued-recall test over all of the pairs that had been presented. Each cue was presented one at a time for 9 s, and subjects were asked to type in the corresponding target word. The pairs were blocked by lists, and the lists were tested in the same order as they had occurred in during the initial learning phase. However, subjects were not told that this was the case. After they `had finished the final test, the subjects were thanked and debriefed. The experiment lasted about three and a half to four hours during one long session.

Results

On each test, the computer scored each item as correct if the first three letters of the item were correct. For example, if the pair was “soccer–piano,” the subject received credit if the first three letters typed were “pia.” This type of scoring allowed for slight misspellings and singular/plural forms of the target words to be considered correct. We also looked at the results using perfect spelling as the criterion for a correct response on the test. Using the first-three-letter scoring method and the perfect spelling method yielded the same pattern of results.

Figure 2 depicts cued-recall performance as a function of the number of study trials for both the standard and repeated-study conditions, with curves plotted in the same way as in Experiment 1. The curves indicate that the standard condition produced better performance than did the repeated-study condition. We performed a 2 (condition:

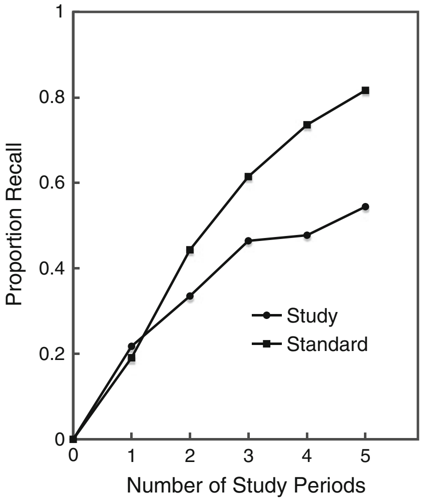

Fig. 2 Standard and pure-study cued-recall learning curves in Experiment 2

standard vs. repeated study) × 5 (study trials: one, two, three, four, or five) ANOVA on the initial cued-recall data. The results from Experiment 2 paralleled those from Experiment 1. Of course, performance increased as a function of the number of study trials [$F(4, 92) = 80.25$, $\eta_p^2 = .78$]. Performance in the standard condition was superior to that in the repeated-study condition [$F(1, 23) = 24.88$, $\eta_p^2 = .52$], indicating again that practicing retrieval enhanced recall overall. Importantly, the Trial × Condition interaction was also significant [$F(1, 23) = 14.26$, $\eta_p^2 = .38$], indicating that learning increased at a faster rate in the standard condition. As we noted in the Experiment 1 results, performance on the first test in both conditions was essentially the same. However, the data after only one study period were collected for both conditions in Experiment 2. We wanted to ensure that the outcome (specifically, the interaction) was not only driven by similar performance after just one study trial. Therefore, we also submitted the initial data to a 2 (condition) × 4 (trials: two, three, four, or five) ANOVA. All *F*s were still significant; the interaction *F* was 6.59, $\eta_p^2 = .57$.

Proportions correct on both the initial and final tests for each of the six conditions are reported in Table 2. The initial recall data in the top row are the same as those in Fig. 2, and the proportions of recall on the final test in the second row are the new data. Recall that in our procedure, subjects used a different list of word pairs for each condition used in creating the pure-study learning curve, as well as one additional list for the standard condition. Examining the second row, one can see that performance increased with the number of study opportunities in the repeated-study condition, but that the alternating study–test (standard) condition produced the best overall recall by a rather wide margin (.77 in the standard condition vs. .52 in the repeated-study condition with five study trials, a 25% gain with numbers of study trials equated).

The data in the second row of Table 2 were submitted to a one-way ANOVA that showed an effect of condition [$F(5, 115) = 47.50$, $\eta_p^2 = .67$]. Pairwise comparisons (Bonferroni corrected to the .05 level) indicated that final performance was significantly greater in the standard condition ($M = .77$) than in all of the repeated-study conditions. In addition, studying once and studying twice in the repeated-study conditions were significantly different from each other and from all other conditions. No other pairwise comparisons were significant.

Also shown in Table 2 are the proportions forgotten between the initial tests and the final tests for the repeated-study conditions, as well as from the final test during learning to the final test in the experiment for the standard condition. Forgetting was measured as the difference between the initial and the final test in each case, divided by the initial score to create a proportional measure of forgetting (Loftus, 1985). An examination of the third row of Table 2 indicates that the more study trials the subject experienced, the less forgetting occurred from the initial learning phase to the final test in general. Additionally, when comparing the Study 5 and standard conditions, it appears that similar rates of forgetting existed and that forgetting was negligible under both conditions. To summarize, when subjects learned using the standard procedure, they were able to learn more, and proportionally, they were able to retain the information, yielding greater performance on the final test overall.

Discussion

The results of Experiment 2 showed the same pattern as in Experiment 1: The standard study–test learning procedure produced greater recall in paired-associate learning than did the pure-study procedure. Thus, the standard study–test procedure produces more learning than does one composed only of study trials. The shapes of the learning curves for paired-associate learning in Fig. 2 are different from those in free recall shown in Fig. 1, with the paired-associate learning functions appearing more nearly linear. Of course, this is partly due to the fact that fewer trials were given in Experiment 2, so subjects had not reached asymptote. However, the fact that the learning curves appear to differ for free recall and paired-associate learning has been known for many years. Recently, Leibowitz, Baum, Enden, and Karniel (2010) argued that not all learning curves are exponential in shape and that a sigmoid function (with initially increasing and then decreasing improvement) may be more representative of learning in certain cases. This issue lies outside the scope of the present study.

Two objections can be raised about the results of Experiments 1 and 2. First, rather than testing helping performance in the standard condition relative to the pure-study condition, perhaps the act of playing Pac-Man in the pure-study condition caused interference. By this logic, testing did not

Table 2 Recall performance on the initial test and on the final test, as well as forgetting, in Experiment 2

	Study 1	Study 2	Study 3	Study 4	Study 5	Standard
Initial Prop. Recalled	.22	.34	.46	.48	.54	.82
Final Prop. Recalled	.18	.29	.43	.44	.52	.77
Forgetting	.18	.15	.07	.08	.04	.06

facilitate performance in the standard condition; rather, playing Pac-Man in the pure-study condition caused interference. We regard this possibility as unlikely, because prior experiments (e.g., Roediger & Payne, 1982) have shown little or no interference when a distractor task employs stimuli different from the study events. However, these and other prior experiments did not use word lists and Pac-Man. In addition, the amount of interference in memory from an intervening task depends on the capacity consumed by that intervening task (Crowder, 1967; Roediger & Crowder, 1975; Roediger et al., 1977), and it may be that Pac-Man is particularly demanding. The second possible objection is that the test trials in the standard condition provided subjects with additional exposure to the material, relative to the pure-study condition. Perhaps merely this extra exposure, and not testing per se, produced the advantage in the standard (study–test) learning condition. Experiment 3 was designed to examine both of these concerns.

Experiment 3

The previous two experiments show clearly that testing enhances learning in the standard study–test paradigm relative to a pure-study condition. However, unlike prior research (Karpicke & Roediger, 2007; Tulving, 1967), in our designs the total amounts of exposure time were confounded between conditions. Examination of Table 1 should make the problem clear: The standard condition involves tests that the pure-study conditions do not, and thus the standard condition permits more exposure to the material. This problem did not exist in prior research, because those experiments held the total number of study or test experiences constant and varied the mix of the two (e.g., in Tulving's, 1967, experiment, 24 total study or test trials were given in three different conditions: SSST, STST, and STTT). A critic of our approach (exemplified in Table 1) could argue that the reason for better performance in the standard condition relative to the repeated-study conditions was that the tests simply permitted more time on task or more exposure to the materials. The hypothesis that testing effects are merely due to additional exposure of material has been examined, and virtually always has been shown not to be the reason that testing improves retention (see Roediger & Karpicke, 2006a, pp. 197–198); testing has much greater power on later retention than does restudying information. Still, we need to determine whether differential exposure created by the test trials in the standard condition might be responsible for the effects observed in the first two experiments.

To meet this end, in Experiment 3 we had subjects learn three lists of words using three different learning conditions, using the same general free-recall procedure used in Experiment 1. In the standard condition, the subjects alternated between study and test trials, as is usually the case. In a second condition, we replaced the test trials in the standard condition with filler trials, as in the repeated-study conditions of Experiments 1 and 2. Of course, these first two conditions served to replicate prior work. In a third learning condition, we replaced the test trials in the standard condition (or the Pac-Man episodes in the pure-study condition) with additional study trials. In this third condition, the subjects studied the list twice in a row. If all testing does is to permit reexposure to information, in the third condition recall should be greater than in the standard study–test condition. The reason is that in the study–test condition, subjects were "reexposed" only to those words that they could recall, whereas in the new study–study condition in Experiment 3, the subjects were exposed to 100% of the items a second time (during the same time period during which subjects in the standard condition were being tested on items). Thus, if all testing does is to permit reexposure to material, the results in Experiment 3 should lopsidedly favor the new pure-study condition relative to the standard condition. In addition, this new condition eliminates Pac-Man as part of the control procedure, and thus if we still found an advantage of the standard condition over the new condition, the argument that the advantage of recall in the standard condition in Experiments 1 and 2 was due to Pac-Man causing interference would also be eliminated.

After learning lists in the three conditions, the subjects then completed a final free-recall test over all three lists to assess retention after learning with each procedure. By comparing final-recall performance in the condition with massed study trials to that after learning in the other two conditions, we could determine whether the advantage of the standard study–test condition in our first two experiments was due to differential exposure created by the tests.

We also manipulated retention interval in Experiment 3, such that one group of subjects completed the final free-recall test at the end of the learning session, whereas others completed the final test after 2 days. In comparing the study–test procedure to a condition with several study trials, Zaromb and Roediger (2010) showed greater benefits of testing after a delay, so it seemed worthwhile to assess retention at two different points in time.

Method

Subjects and materials A group of 60 subjects was recruited from the Washington University in St. Louis human subject pool and participated in exchange for partial course credit or payment. Seven of the subjects were removed and replaced because they did not follow directions; in particular, they recalled words from the wrong list during test trials. This problem did not occur in the earlier experiments, and it occurred here because the one list that was never tested tended to intrude into the next list. Szpunar, McDermott, and Roediger (2008) showed that a test trial tends to reduce or even eliminate proactive interference in free recall. None of the subjects had participated in the prior experiments, and three of the lists from Experiment 1 were used for Experiment 3.

Design A 3 (learning condition) × 2 (retention interval) mixed-factorial design was used. Learning condition (standard, pure-study spaced, and pure-study massed) was manipulated within subjects, whereas retention interval (immediate or delayed) was manipulated between subjects. For two learning conditions, the number of study and/or test trials was held constant at 8 whereas in the other condition (pure study-spaced) there were only four study trials. In the standard condition, subjects alternated between studying and recalling the list of words, for a total of four study–test trials. In the pure-study-spaced condition, subjects alternated between studying the list of words and completing a filler task (playing Pac-Man), for four study trials interleaved with four filler trials. In the pure-study-massed condition, subjects studied the list of words for eight consecutive study trials (i.e., study trials replaced test trials in the standard condition and the filler task in the pure-study-spaced condition). We refer to this as the *pure-study-massed condition*, in that the list was repeated eight times with minimal breaks between presentations, although of course presentation of the individual words was spaced, because the lists were presented in a different random order on each trial. As in Experiment 1, list order was held constant, and the order in which the subjects learned using each of the three learning conditions was fully counterbalanced. Altogether, the list was presented four times in the pure-study-spaced condition and eight times in the pure-study-massed condition; in the standard condition, it was presented four times and recalled four times.

Procedure As in the previous experiments, the subjects learned one list for each of the three within-subjects conditions described above. The study, test, and filler trials were the same as in Experiment 1. However, in Experiment 3 we added a warning to the "Begin Study" prompt, reminding subjects to pay attention to the presentation of each item, and they needed to press the Enter key to advance to the first study trial. We did this to try to ensure that subjects were paying attention during learning, especially during the pure-study-massed condition. We also added a warning to the "Begin Test" prompt, reminding subjects to only recall words from the most recent study trials (i.e., reminding subjects not to think back to previous lists). We did this because subjects were not able to take a test after learning each list as they had in the first two experiments, and this could cause subjects to experience more proactive interference from previous lists (Szpunar, McDermott, & Roediger, 2007, 2008). As we noted, this instruction was not completely successful, and some subjects who did not follow the directions had to be eliminated and replaced.

After subjects learned each of the three lists, they completed another filler task (they played Tetris) for 15 min. After this task, subjects in the immediate condition completed a final free-recall test. During this test, they were given 12 min to recall as many words from the learning phase as they could, from all lists. The subjects were warned not to guess during this test, and they were encouraged to try to recall as many words as they could from all three lists. The subjects in the delayed condition were sent home and asked to return to the lab 2 days later. When they returned, they completed the same final free-recall test as those in the immediate condition. After their final test, all of the subjects were debriefed and thanked for their time.

Results

Scoring was completed in the same way as in Experiment 1. Figure 3 depicts initial recall performance across each of the four test periods during the standard learning condition for subjects in both the immediate and delayed conditions. Of course, no variable had been manipulated at this point, so the curves were expected to be highly similar. A 4 (test period: one, two, three, and four) × 2 (retention interval: immediate vs. delayed) ANOVA revealed only a significant effect of test period [$F(3, 174) = 166.00$, $\eta_p^2 = .74$], indicating that subjects improved their performance across study and test trials during the standard learning condition. We found no main effect of retention interval during initial learning [$F(1, 58) = 1.52$, $\eta_p^2 = .03$], nor any interaction

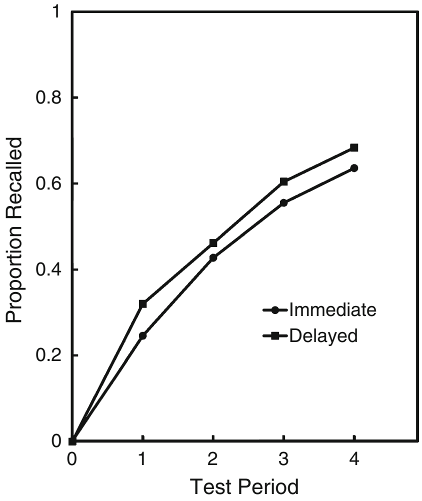

Fig. 3 Initial learning during the standard condition in Experiment 3. The immediate and delayed test conditions were not instantiated until after learning, so no difference was expected between the two in learning

($F < 1$), indicating that the performance in the two groups was roughly equivalent, as expected. On average, subjects in the standard condition recalled 49% of the items on their four tests, whereas subjects in the massed condition were presented with 100% of the items during the same time period. Thus, the number of presentations favored the massed-study relative to the standard condition by a 2:1 ratio. Would this difference overcome the testing effect?

The answer is no, as can be seen in Fig. 4, which depicts performance on the final free-recall test. Lists learned in the standard (study–test) condition were recalled better on both the immediate and delayed final free-recall tests than were lists in either the pure-study-spaced condition (with four study presentations) or the pure-study-massed condition (with eight presentations). Thus, the tests in the standard condition clearly had a greater mnemonic benefit than simply providing more exposure of the list or more time on task. Despite the greater number of item presentations in the massed condition, with eight study presentations, recall was better in the standard condition with four study trials and four test trials. In addition, eliminating Pac-Man from a pure-study condition did not eliminate the facilitating effect of testing (even when Pac-Man was replaced by additional study trials).

These conclusions were confirmed with a 2 (retention interval: immediate or delayed)×3 (initial learning condition: standard, massed, or spaced) ANOVA on the final free-recall data. As expected, a main effect of retention interval [$F(1, 58) = 10.82$, $\eta_p^2 = .16$] was present, indicating that forgetting occurred during the 2-day delay. There was also an effect of initial learning condition [$F(2, 116) = 54.51$, $\eta_p^2 = .48$], indicating an overall effect of the three conditions. Post-hoc analyses (Bonferroni corrected to the .05 level) indicated that all three learning conditions were significantly different from one another. The standard learning condition ($M = .61$) resulted in the greatest final free recall, the spaced learning condition led to the lowest recall performance ($M = .32$), and the massed learning condition led to performance between the two ($M = .44$). Thus, eight massed presentations led to somewhat greater recall than did four spaced presentations. The interaction was not significant ($F < 1$): The relationship among the initial learning conditions was the same, whether retention was assessed immediately or after a 2-day delay.

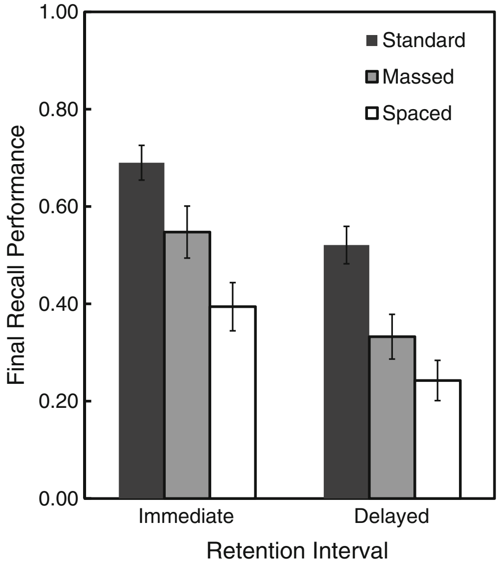

Fig. 4 Final free-recall performance in Experiment 3 as a function of the three conditions: standard, pure-study massed (eight list presentations), and pure-study spaced (four list presentations)

As in Experiment 2, we calculated forgetting that occurred from the immediate retention test to the delayed retention test. Even though retention interval was manipulated between subjects, we would expect that if subjects in the delayed condition had taken an immediate free-recall test, they would have performed similarly to those in the immediate condition. To assess forgetting, for each of the three initial learning conditions, we subtracted the proportion correct for the delayed group from the proportion correct for the immediate group and divided by the immediate group's score to create a proportional measure of forgetting (see Loftus, 1985). The proportions of forgetting were similar when the material was learned with massed study trials (.39) and with spaced study trials (.38). However, it seems that learning by alternating between study and test trials slowed the rate of forgetting (.25; see also Carpenter, Pashler, Wixted, & Vul, 2008; Roediger & Karpicke, 2006b). We could not perform statistical analyses of these results because of the nature of the design, in which retention interval was manipulated between subjects.

General discussion

In the first two experiments, using free recall and paired-associate learning, we showed that learning curves produced without cumulative testing grew less rapidly than standard learning curves with cumulative testing. Contrary to the implications of prior research (Tulving, 1967; see also Donaldson, 1971; Karpicke & Roediger, 2007; among others), test trials do enhance learning in standard study–test procedures. This may be due to direct effects of testing or indirect effects of the tests potentiating learning on additional study trials (Arnold & McDermott, 2012; Izawa, 1971). Although theories of learning have implicitly assumed that learning only occurs during the study trials in multitrial procedures, our results show that this assumption is incorrect. A learning curve based solely on study trials produces poorer performance than does the standard study–test procedure. For example, in Experiment 2, when the numbers of study trials were matched between the two procedures (i.e., comparing the Study 5 condition to the fifth test trial of the standard condition), a 25% advantage in paired-associate recall favored the standard condition. This increase must be attributed to the test trials inserted between each study trial, in this case. In Experiment 3, we investigated whether differential exposure on test trials could account for the advantage of the standard conditions over the repeated-study conditions in Experiments 1 and 2. The answer is no: Performance after learning a list with eight massed study trials was inferior to that in the standard condition of four study and four test trials, despite the fact that the sheer number of item presentations was much greater in the former than in the latter condition. This result held after both a relatively short retention interval (within an experimental session) and a much longer retention interval (2 days). The inferior performance in the repeated-study-massed condition relative to the standard condition also shows that the advantage of the standard conditions to the repeated-study conditions in Experiments 1 and 2 could not be attributed to the Pac-Man filler task, because that task was not used in the repeated-study-massed condition.

One interesting point that emerges from comparing our experiments to previous research (e.g., Tulving, 1967) is how few test trials it takes to bring learning to the same level as in the standard study–test condition. As noted above, in Tulving's (1967) experiment, only six test trials produced a learning curve that was more or less equivalent to one in which 18 trials were used, with the number of study and/or test trials held constant (see also Karpicke & Roediger, 2007). Yet in the present Experiment 1, in which we also used free recall, reducing the number of test trials to zero greatly affected performance relative to having eight test trials, and repeated exposure can account for only a small part of this difference.

Why do test trials boost learning? We suspect that several factors are at work and that the answer may be different for free-recall and paired-associate learning procedures. One common factor between these methods is the potentiating effect that test trials can have on future study trials (Arnold & McDermott, 2012; Izawa, 1971). Although the reasons for test potentiation are not yet well understood, relevant evidence exists for both free recall and paired-associate learning. Battig, Allen, and Jensen (1965) examined recall protocols during multitrial free-recall experiments and discovered that the first items recalled tended to be items that had not been recalled on the prior trial. One reasonable interpretation of this finding is that when subjects restudy a list after a test trial, they focus their attention and encoding efforts on items they recognize as not having been recalled on the previous trial. Hence, they recall these items first. If so, this may be one way that potentiation works in free recall. During paired-associate learning, Pyc and Rawson (2010) suggested that subjects often develop mediators to link stimuli and responses, and that testing causes them to realize when mediators are ineffective and to try new ones. These researchers reported evidence for this mediator effectiveness hypothesis, as did Carpenter (2011).

In the case of free recall, additional processes are probably aided by retrieval. Forming an organizational scheme is critical for free recall, and evidence has shown that subjects organize their recall either using a structure built into the material (Bousfield, 1953) or using idiosyncratic subjective organization (Tulving, 1962; see also Mandler, 1967). Organized recall probably arises because subjects create a retrieval plan or schema to guide recall. Zaromb and Roediger (2010) showed that test trials led to greater categorical organization (organizing randomly presented items into categories) and subjective organization (the tendency to keep the recall sequences the same from trial to trial) when the learning phase included increasng numbers of test trials. Their findings were consistent with the hypothesis that testing permits subjects to form schemas for retrieval (see also Zaromb, 2011). Study trials probably do not afford the opportunity for this type of processing—the creation of retrieval plans or schemas—which is required for success in free recall; at least, they do not afford the opportunity as well as test trials do.

Paired-associate learning is different. After studying pairs (A–B), subjects are given one member of the pair and must recall the other, so the associative bond is critical—the process of being able to recall B after A (McGuire, 1961). Test trials give subjects practice at this task that study trials do not, and consequently, test trials have much greater impact on long-term retention than do study trials (Karpicke, 2009; Karpicke & Roediger, 2008). After studying an A–B pair, testing increases not only recall of B when given A, but also the reverse, improved recall of A when given B as a cue (Carpenter, Pashler, & Vul, 2006). Of course, testing may help learning in other ways, too.

At a more general level, however, testing may improve learning in both free recall and paired-associate learning by enhancing relational processing (Hunt & McDaniel, 1993). Furthermore, the benefits of testing generally follow the principle of transfer-appropriate processing (Bransford, Franks, Morris, & Stein, 1979; Roediger, Gallo, & Geraci, 2002), in that testing (relative to studying) permits subjects to practice the skill that will be needed on the criterial test. In the case of free recall, the skill is using a retrieval plan or organizational scheme to guide recall of the list with minimal cues; in the case of paired-associate learning, the skill is the ability to retrieve one member of the pair when given the other. Thus, at this broader level, the testing effect conforms to these more general principles.

Author note This research was supported by a Collaborative Activity Grant from the James S. McDonnell Foundation. We thank Ben Hoener and Kelly Young for assistance with data collection, and Andy DeSoto, Robert Greene, Alice Healy, Hal Pashler, and Adam Putnam for helpful comments.

References

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1063–1087. doi:10.1037/0278-7393.20.5.1063

Arnold, K. M., & McDermott, K. B. (2012). *Test-potentiated learning: Distinguishing between direct and indirect effects of tests*. Manuscript submitted for publication.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459. doi:10.3758/BF03193014

Battig, W. F., Allen, M., & Jensen, A. R. (1965). Priority of free recall of newly learned items. *Journal of Verbal Learning and Verbal Behavior*, *4*, 175–179.

Bernbach, H. A. (1970). A multiple-copy model for post-perceptual processing. In D. A. Norman (Ed.), *Models of human memory* (pp. 103–116). New York, NY: Academic Press.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Erlbaum.

Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology*, *49*, 229–240.

Bransford, J. D., Franks, J. J., Morris, C. D., & Stein, B. S. (1979). Some general constraints on learning and memory research. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 331–354). Hillsdale, NJ: Erlbaum.

Bregman, A. S., & Wiener, J. R. (1970). Effects of test trials in paired-associate and free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *9*, 689–698.

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1547–1552. doi:10.1037/a0024140

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, *13*, 826–830.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition*, *36*, 438–448.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*, 633–642. doi:10.3758/BF03202713

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review of quantitative synthesis. *Psychological Bulletin*, *132*, 354–380. doi:10.1037/0033-2909.132.3.354

Crowder, R. G. (1967). Short-term memory for words with a perceptual–motor interpolated activity. *Journal of Verbal Learning and Verbal Behavior*, *6*, 753–761.

Donaldson, W. (1971). Output effects in multitrial free recall. *Journal of Verbal Learning and Verbal Behavior*, *10*, 577–585.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). New York, NY: Dover. (Original work published in 1885)

Estes, W. K. (1960). Learning theory and the new “mental chemistry.” *Psychological Review*, *67*, 207–223. doi:10.1037/h0041624

Estes, W. K. (1988). Human learning and memory. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & D. R. Luce (Eds.), *Stevens' handbook of experimental psychology: Vol. 2. Learning and cognition* (2nd ed., pp. 351–415). New York, NY: Wiley.

Estes, W. K., Hopkins, B. L., & Crothers, E. J. (1960). All-or-none and conservation effects in the learning and retention of paired associates. *Journal of Experimental Psychology*, *60*, 329–339.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, *6*.

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

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 801–812.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*, 185–207. doi:10.3758/BF03212979

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

Hovland, C. I. (1951). Human learning and retention. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 613–689). New York, NY: Wiley.

Hull, C. L. (1952). *A behavior system: An introduction to behavior theory concerning the individual organism.* New Haven, CT: Yale University Press.

Hunt, R. R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory and Language*, *32*, 421–445.

Izawa, C. (1967). Function of test trials in paired-associate learning. *Journal of Experimental Psychology*, *75*, 194–209.

Izawa, C. (1971). The test trial potentiating model. *Journal of Mathematical Psychology*, *8*, 200–224. doi:10.1016/0022-2496(71)90012-5

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*, 469–486.

Karpicke, J. D., & Roediger, H. L., III. (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, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966–968. doi:10.1126/science.1152408

Lachman, R., & Laughery, K. R. (1968). Is a test trial a training trial in free recall learning? *Journal of Experimental Psychology*, *76*, 40–50.

Leibowitz, N., Baum, B., Enden, G., & Karniel, A. (2010). The exponential learning equation as a function of successful trials results in sigmoid performance. *Journal of Mathematical Psychology*, *54*, 338–340.

Loftus, G. R. (1985). Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 397–406. doi:10.1037/0278-7393.11.2.397

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*, 203–208. doi:10.3758/BF03204766

Mandler, G. (1967). Organization and memory. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 1, pp. 327–372). New York: Academic Press.

Mazur, J. E., & Hastie, R. (1978). Learning as accumulation: A reexamination of the learning curve. *Psychological Bulletin*, *85*, 1256–1274.

McGeoch, J. A. (1942). *The psychology of human learning: An introduction.* New York, NY: Longmans, Green.

McGuire, W. J. (1961). A multiprocess model for paired-associate learning. *Journal of Experimental Psychology*, *62*, 335–347.

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior*, *9*, 596–606.

Miller, G. A., & McGill, W. J. (1952). A statistical description of verbal learning. *Psychometrika*, *17*, 369–396.

Patterson, K. E. (1972). Some characteristics of retrieval limitation in long-term memory. *Journal of Verbal Learning and Verbal Behavior*, *11*, 685–691.

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

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*, 335. doi:10.1126/science.1191465

Rock, I. (1957). The role of repetition in associative learning. *American Journal of Psychology*, *70*, 186–193.

Rock, I., & Heimer, W. (1959). Further evidence of one-trial associative learning. *American Journal of Psychology*, *72*, 1–16.

Roediger, H. L., III. (1974). Inhibiting effects of recall. *Memory & Cognition*, *2*, 261–269. doi:10.3758/BF03208993

Roediger, H. L., III, & Arnold, K. M. (2012). The one-trial learning controversy and its aftermath: Remembering Rock (1957). *American Journal of Psychology*, *125*, 127–143. doi:10.5406/amerjpsyc.125.2.0127

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

Roediger, H. L., III, & Crowder, R. G. (1975). The spacing of lists in free recall. *Journal of Verbal Learning and Verbal Behavior*, *14*, 590–602.

Roediger, H. L., III, Gallo, D. A., & Geraci, L. (2002). Processing approaches to cognition: The impetus from the levels of processing framework. *Memory*, *10*, 319–332.

Roediger, H. L., III, & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210. doi:10.1111/j.1745-6916.2006.00012.x

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

Roediger, H. L., III, Knight, J. L., & Kantowitz, B. H. (1977). Inferring decay in short-term memory: The issue of capacity. *Memory & Cognition*, *5*, 167–176.

Roediger, H. L., III, & Payne, D. G. (1982). Hypermnesia: The role of repeated testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*, 66–72.

Roediger, H. L., III, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. Mestre & B. H. Ross (Eds.), *The psychology of learning and motivation: Cognition in education* (pp. 1–36). Amsterdam, The Netherlands: Elsevier.

Rosner, S. R. (1970). The effects of presentation and recall trials on organization in multitrial free recall. *Journal of Verbal Learning and Verbal Behavior*, *9*, 69–74.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L., III. (2007). Expectation of a final cumulative test enhances long-term retention. *Memory & Cognition*, *35*, 1007–1013. doi:10.3758/BF03193473

Szpunar, K. K., McDermott, K. B., & Roediger, H. L., III. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1392–1399. doi:10.1037/a0013082

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

Tulving, E. (1962). Subjective organization in free recall of “unrelated” words. *Psychological Review*, *69*, 344–354. doi:10.1037/h0043150

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.

Underwood, B. J., & Keppel, G. (1962). One trial learning? *Journal of Verbal Learning and Verbal Behavior*, *1*, 1–13.

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11*, 571–580.

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

Zaromb, F. M. (2011). Organizational processes contribute to the testing effect in free recall (Doctoral dissertation, Washington University in St. Louis, 2011). *Dissertation Abstracts International*, *71*, 5818.

Zaromb, F. M., & Roediger, H. L., III. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition*, *38*, 995–1008. doi:10.3758/MC.38.8.995