Journal of Experimental Psychology:
Learning, Memory, and Cognition
2007, Vol. 33, No. 4, 704–719

0278-7393/07/$12.00 DOI: 10.1037/0278-7393.33.4.704

Expanding Retrieval Practice Promotes Short-Term Retention, but Equally Spaced Retrieval Enhances Long-Term Retention

Jeffrey D. Karpicke and Henry L. Roediger III
Washington University in St. Louis

Expanding retrieval practice (T. K. Landauer & R. A. Bjork, 1978) is regarded as a superior technique for promoting long-term retention relative to equally spaced retrieval practice. In Experiments 1 and 2, the authors found that expanding retrieval practice of vocabulary word pairs produced short-term benefits 10 min after learning, conceptually replicating Landauer and Bjork's results. However, equally spaced retrieval produced superior retention 2 days later. This pattern occurred both with and without feedback after test trials. In Experiment 3, the 1st test occurred immediately or after a brief delay, and repeated tests were expanding or equally spaced. Delaying the first test improved long-term retention, regardless of how the repeated tests were spaced. The important factor for promoting long-term retention is delaying initial retrieval to make it more difficult, as is done in equally spaced retrieval but not in expanding retrieval. Expanding the interval between repeated tests had little effect on long-term retention in 3 experiments.

Keywords: testing effect, spacing effect, retrieval practice, learning, retention

Expanding retrieval practice is often advocated as a method of improving long-term retention, especially for the purposes of bolstering student learning and enhancing memory in older adults and in memory-impaired populations (Bjork, 1988; Camp, Bird, & Cherry, 2000; Cull, Shaughnessy, & Zechmeister, 1996; Landauer & Bjork, 1978; Schmidt & Bjork, 1992). The technique involves attempting to retrieve an item immediately after it has been studied (an immediate first test) and then gradually increasing the spacing interval between successive retrieval attempts. This expanding retrieval procedure is intended to ensure a high level of retrieval success on the first test and to increase the difficulty of retrieval attempts on subsequent repeated tests. Gradually increasing the spacing of repeated tests is considered a shaping procedure for long-term retention, because expanding the schedule of repeated tests is thought to increasingly approximate the conditions of a delayed final test (see Schmidt & Bjork, 1992). Although expanding retrieval is widely believed to be an effective technique to enhance learning, previous research is inconsistent about whether expanding retrieval works. We carried out the present experiments to determine whether expanding retrieval represents a superior form of spaced retrieval practice for promoting long-term retention.

The Testing Effect and Expanding Retrieval Practice

A considerable amount of research has shown that taking a memory test over some material can improve long-term retention, relative to repeatedly studying the material. This phenomenon is known as the testing effect (e.g., Bjork, 1975; Carrier & Pashler, 1992; Chan, McDermott, & Roediger, 2006; McDaniel & Masson, 1985; Roediger & Karpicke, 2006a, 2006b; Wheeler & Roediger, 1992). The testing effect is particularly surprising considering that it occurs even when subjects are not given feedback about their test performance, because when tests involve recall subjects can only reexperience whatever they are able to produce on the test. For example, Roediger and Karpicke (2006b) had subjects read a prose passage and then either restudy the passage or recall as much of it as they could on a free recall test. Even though subjects could recall about 70% of the passage on the initial test, taking the test led to better final recall 2 days or 1 week later than studying the entire passage again. Because the testing effect occurs even when feedback is not provided after the test, and because testing enhances learning more than additional studying, the act of retrieval when taking a test is critical in promoting long-term retention (Karpicke & Roediger, in press; Roediger & Karpicke, 2006a).

Just as testing produces positive effects on later retention, spacing repetitions across time or other intervening events also enhances retention. When the same item is studied twice within a list, later retention of the material generally increases as a negatively accelerated function of the spacing or lag between repetitions (see Madigan, 1969; Melton, 1970). Spacing effects are found even when the spacing interval between study periods is very long and especially after long retention intervals (see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). However, the general rule that greater spacing intervals produce increasing benefits for retention may not remain true when the second repetition is a test, because increasing the spacing between a study period and a test will also decrease the likelihood of recall on that test. In other words, forgetting during

Jeffrey D. Karpicke and Henry L. Roediger III, Department of Psychology, Washington University in St. Louis.

This research was supported by grants from the Institute for Education Sciences and the James S. McDonnell Foundation. We thank Julie Evans and Elizabeth Rhodes for assistance with data collection. We also thank Dave Balota and Jessica Logan for comments.

Correspondence concerning this article should be addressed to Jeffrey D. Karpicke or to Henry L. Roediger III, Department of Psychology, Washington University, Campus Box 1125, One Brookings Drive, St. Louis, MO 63130-4899. E-mail: karpicke@wustl.edu or roediger@artsci.wustl.edu

the spacing interval may counteract any benefits due to spacing the test. Expanding retrieval practice is intended to remedy this particular problem by incorporating a test immediately after studying, thereby minimizing forgetting on the first test, and then gradually increasing the spacing of repeated tests.

The idea behind expanding retrieval practice is intuitive, and references to similar procedures can be found in early literature on human learning. For example, in an early textbook on educational psychology, Starch (1927) wrote, "Since the rate of forgetting is very rapid at first and more gradual later on, it probably would be highly advantageous to have relearning of a given material come very frequently at first and more rarely later on" (p. 156).

In an important article, Landauer and Bjork (1978) introduced the idea of expanding retrieval practice to contemporary researchers. In their experiments, subjects studied pairs of items (e.g., names paired with faces, or first names paired with last names) and then took three or four cued recall tests in a continuous paired associates task (see Glenberg, 1976; Peterson, Wampler, Kirkpatrick, & Saltzman, 1963). Landauer and Bjork created expanding and equally spaced conditions by varying the number of trials that occurred between the study trial and the repeated tests. For example, in one of their expanding retrieval conditions, the first test occurred one trial after the study trial, a second test occurred after four more trials, and a third test occurred after 10 more trials. This condition was denoted 1–4–10 to indicate the number of intervening trials between study or test trials. Performance in this expanding retrieval condition was compared with an equally spaced condition in which five trials occurred between the study trial and the subsequent test trials (denoted 5–5–5). Although the distribution of the tests differed in these two conditions, the average spacing interval was five trials in both conditions. Landauer and Bjork found that on a final test 30 min after the learning phase, expanding retrieval practice produced about a 10% advantage over equally spaced retrieval practice.

Bjork (1994, 1999) has argued that expanding retrieval works because it introduces desirable difficulties during learning that improve later retention. The idea behind desirable difficulties is that techniques requiring more effortful processing on the part of the learner may sometimes slow initial learning, relative to other techniques that promote less effortful processing, but the more difficult conditions will ultimately lead to greater long-term retention (see also McDaniel & Einstein, 2005). For example, spaced practice represents a desirable difficulty, relative to massed practice, because massed practice often leads to better performance immediately after learning, but spacing leads to better long-term retention on delayed tests (see Bahrick, 1979; Balota, Duchek, & Paullin, 1989; Glenberg, 1976; Peterson et al., 1963). The testing effect can also be considered a desirable difficulty: Although repeated studying leads to better performance than repeated testing on criterial tests immediately after learning, testing leads to better long-term retention than repeated studying (see Hogan & Kintsch, 1971; Roediger & Karpicke, 2006b; Thompson, Wenger, & Bartling, 1978; Wheeler, Ewers, & Buonnano, 2003). Likewise, expanding retrieval practice is thought to create difficulties during learning that have positive effects on long-term retention. Gradually increasing the interval between repeated tests is assumed to make each successive retrieval attempt increasingly more difficult, relative to the constant levels of retrieval difficulty involved in equally spaced schedules. The increased retrieval difficulty involved on the expanding tests, coupled with high levels of retrieval success due to an immediate first test, are thought to produce optimal long-term retention (Landauer & Bjork, 1978; see too Bjork, 1988, 1994, 1999).

Following Landauer and Bjork (1978), many authors have argued that expanding retrieval practice represents a powerful means of improving long-term retention. Rea and Modigliani (1985) and Cull et al. (1996) argued that expanding retrieval should be applied in educational settings as a way to improve student learning. Schacter, Rich, and Stampp (1985) used expanding retrieval to improve memory in amnesic patients. Camp and his colleagues (e.g., Camp, 2006; Camp et al., 2000) have used expanding retrieval as a memory rehabilitation technique for people with dementia (but see Balota, Duchek, Sergent-Marshall, & Roediger, 2006). Considering the widespread belief in the utility of expanding retrieval, it is surprising that there is not a larger base of research with consistent evidence showing expanding retrieval practice to be the superior spaced practice technique for improving long-term retention.

Prior Research on Expanding Retrieval

Initial investigations of expanding retrieval appeared 7 years after Landauer and Bjork's (1978) original chapter, but not all have addressed the issue of whether expanding retrieval represents a superior form of spaced practice relative to equal spacing (see Balota, Duchek, & Logan, 2007, for review). Rea and Modigliani (1985) had children learn spelling words or multiplication problems and found benefits of expanding retrieval practice over massed practice. However, they did not compare their expanding condition with an equally spaced condition, so the advantage of expanding retrieval in this case could simply be a spacing effect rather than a benefit of expanding retrieval, per se. Research by Schacter et al. (1985) and Camp and colleagues (Camp, 2006; Camp et al., 2000) has also observed benefits of expanding retrieval relative to pretest baseline measures of memory performance with patient populations. Again, in these studies, it is not clear whether repeated testing, or spaced testing, or the particular schedule of spaced testing is responsible for positive effects on retention.

Shaughnessy and Zechmeister (1992) were the first after Landauer and Bjork (1978) to make the critical comparison of an expanding schedule of retrieval practice with an equally spaced schedule. They had subjects learn word pairs in a continuous paired associate task (like the one used by Landuaer & Bjork, 1978) and gave a criterial test shortly after learning. They found a 6% advantage of expanding retrieval over equally spaced practice, although this effect was not statistically significant. Cull et al. (1996) carried out a more extensive analysis with five experiments, again using a continuous paired associate task with a final test immediately after the learning phase. However, they obtained significant positive effects of expanding retrieval over equally spaced practice in only two of the five experiments.

More recent experiments have also not fared well in obtaining positive effects of expanding retrieval practice. Carpenter and DeLosh (2005) compared expanding versus equally spaced retrieval conditions in two experiments, using materials and spacing conditions similar to those of Cull et al. (1996), but they did not obtain positive benefits of expanding retrieval on a 5-min delayed test in either experiment. Balota et al. (2006) investigated massed, expanding, and equally spaced retrieval practice in three groups of subjects: younger adults, healthy older adults, and older adults with dementia of the Alzheimer's type (DAT). The subjects studied word pairs and took two immediate tests after studying each pair, to ensure that subjects had encoded the items. Then three repeated tests were spaced according to massed, expanding, or equal interval schedules. Balota et al. found that on a criterial test at the end of the learning phase, the spaced practice conditions led to better performance than massed practice in all three groups of subjects, but there were no differences between expanding and equally spaced practice. In some cases equally spaced practice produced a modest benefit over expanding retrieval on the final test. Hochhalter, Overmier, Gasper, Bakke, and Holub (2005) also tested individuals with DAT and also did not observe differences between expanding and equally spaced schedules of retrieval practice.

Interestingly, the results of two studies suggest that advantages of equally spaced practice over expanding retrieval sometimes occur when long-term retention is assessed after a delay (in a separate session days or weeks after learning). Cull (2000) conducted four experiments and used conditions without feedback (following Landauer & Bjork's, 1978, original procedure) and with feedback after each test, using a technique introduced by Carrier and Pashler (1992) to equate the total time involved in test trials and test-plus-feedback trials. Cull did not report any significant positive effects of expanding retrieval, and in some conditions he found that equally spaced retrieval practice led to superior long-term retention after a delay of 3 or 8 days. More recently, Logan and Balota (in press) examined expanding and equally spaced retrieval practice schedules at short and long retention intervals. They found no advantage of expanding retrieval on a final test immediately after learning, but they found an advantage of equal spacing on a final test after a 24-hr delay. (We provide further discussion of Logan & Balota's work in the General Discussion section.)

Given that expanding retrieval is widely believed to be a superior technique for improving long-term retention, it seems surprising that prior research has not shown consistent advantages of expanding retrieval practice over equally spaced practice on criterial tests given shortly after the learning period, as in Landauer and Bjork (1978). Further, it seems even more surprising that equally spaced practice sometimes leads to greater recall than expanding practice on tests given after a day or more delay.

Why might equally spaced retrieval practice produce better long-term retention than expanding retrieval? We propose that the difficulty of the initial retrieval attempt may be critical in determining long-term retention and that the particular schedule of retrieval (equal interval or expanding) may not matter much except for the inherent manipulation of the timing of the initial test in the two schedules. An immediate first test may enhance short-term accessibility during the learning phase, but a somewhat delayed first test (requiring more difficult retrieval) is necessary to enhance long-term retention. By this logic, the reason that Landauer and Bjork (1978) obtained an advantage of expanding retrieval on relatively short-term tests (around 30-min retention intervals) was due to the enhanced short-term accessibility conferred by the relatively immediate first test in the expanding retrieval condition. However, this initial test was too easy to establish a benefit in long-term retention as measured several days later (Cull, 2000). The equal interval schedule, on the other hand, provides a delayed first test, which harms accessibility during learning but enhances processes that support long-term retention. By this account, the critical factor for increasing long-term retention is providing an initial test in which recall is possible but relatively difficult. The optimum range of delay for the test will probably depend on several factors, but we hypothesize that the representation of an event must at least have been cleared from short-term or working memory for a test to promote long-term retention. If retrieval occurs from primary memory, there will probably be little advantage in the long term. Maintenance rehearsal is a form of repeated retrieval from short-term memory and provides little or no benefit to recall (e.g., Craik & Watkins, 1973). Of course, the optimum delay for a test to have a positive effect on long-term retention may depend on the task employed and other factors, as we elaborate in the General Discussion section.

To reiterate, our main point is that the first retrieval attempt must provoke some effort and, we hypothesize, should occur after the item's presentation has been cleared from primary memory. Prior research has shown that the difficulty of initial retrieval is correlated with later retention (Benjamin, Bjork, & Schwartz, 1998; Gardiner, Craik, & Bleasdale, 1973). There is also direct evidence that delaying an initial retrieval attempt enhances performance on a later criterial test (Jacoby, 1978; Modigliani, 1976; Whitten & Bjork, 1977). However, in prior research on expanding versus equally spaced retrieval practice, performance on tests during the learning phase was typically not examined. Indeed, some experiments did not require subjects to make any overt responses during learning (e.g., Cull et al., 1996), preventing any possible analysis of learning phase performance. Thus, it is not clear that delaying the first test makes retrieval more difficult and thereby promotes long-term retention, as predicted by our hypothesis, nor is it clear that expanding the interval between repeated tests gradually increases the difficulty of retrieval on those tests, as predicted by Landauer and Bjork's (1978) theory. Our alternate theory is also based on several assumptions that may be questioned, but our experiments provide relevant evidence.

One goal of the present experiments was to examine performance during the learning phase in expanding and equally spaced retrieval practice conditions. To this end, we used response latency for correct recalls as an index of retrieval difficulty, following several other researchers (Benjamin & Bjork, 1996; Benjamin et al., 1998; Gardiner et al., 1973; Kelley & Lindsay, 1993; Koriat & Ma'ayan, 2005; Matvey, Dunlosky, & Guttentag, 2001; Serra & Dunlosky, 2005). If expanding retrieval makes repeated tests increasingly difficult, this should be reflected in increasingly slow response latencies across repeated tests. Likewise, if delaying the first test in the equally spaced condition increases the difficulty of the initial retrieval attempt, this effect would be reflected in slower response times on the delayed first test, relative to the immediate first test in the expanding retrieval condition. We also provided single test control conditions to examine our assumptions about delayed tests (relative to tests after short delays), providing a greater testing effect on long-term retention even if producing poorer performance during the learning phase itself.

The Present Experiments

In the present experiments, subjects studied vocabulary word pairs and took tests over them spaced according to several different schedules. With the potential educational relevance of expanding retrieval practice in mind, we had subjects learn vocabulary word pairs selected from test preparation books for the Graduate Record Exam (GRE) that college students might try to learn using some type of spaced retrieval practice technique. Experiments 1 and 2 were carried out simultaneously and investigated massed, expanding, and equally spaced retrieval practice conditions, similar to those used in previous experiments (Cull et al., 1996; Landauer & Bjork, 1978; Shaughnessy & Zechmeister, 1992). In Experiment 1, subjects were not given feedback about their responses on tests (following Landauer & Bjork's, 1978, original procedure), but in Experiment 2, subjects were given feedback. Providing feedback after tests in the learning phase should counteract forgetting that would occur on a delayed first test in the equally spaced condition. Thus, any differences in learning phase performance in the expanding and equally spaced conditions may be eliminated by providing feedback, whereas any positive effects of delaying the first test in the equally spaced condition would be preserved. In both experiments, retention was assessed on a final criterial test given either 10 min or 2 days after learning. The 10-min delay is similar to that used by Landauer and Bjork (1978), whereas the 2-day retention interval provides a test of longer term retention. According to our hypothesis, equally spaced practice should lead to better long-term retention than expanding retrieval, because delaying the first test in the equally spaced condition increases retrieval difficulty, which promotes long-term retention.

Experiment 3 was designed to separate the effects of delaying the first test and the schedule of repeated tests. In previous comparisons of expanding and equally spaced practice, these two factors are naturally confounded (but see Carpenter & DeLosh, 2005, discussed below). Expanding retrieval typically involves an immediate first test, whereas equally spaced practice involves a delayed first test. Any differences between expanding and equally spaced retrieval practice could lie in differences in the position of the first test or in the schedule of repeated tests. Experiment 3 separated the effects of these two variables by factorially manipulating the position of the first test and the schedule of repeated tests. According to our hypothesis, delaying the first test should have positive effects on long-term retention. We suspected that control of this variable would negate any further difference between expanding and equal interval schedules of testing.

Experiment 1

In Experiment 1, we examined three spaced testing conditions: Massed (0–0–0), expanding (1–5–9), and equally spaced (5–5–5). In addition, two single-test conditions were included to provide converging evidence about the effects of delaying an initial test on later retention and also to investigate the effects of repeated testing relative to taking a single test. One single-test condition involved an immediate test (after a spacing of one item), and the other condition involved a delayed test (after a spacing of five items). Subjects did not receive feedback after the tests in Experiment 1. Half of the subjects took the final test after a 10-min retention interval, and half took the final test 2 days after the learning phase. We predicted that equally spaced retrieval practice would produce better long-term retention than expanding retrieval on a delayed criterial test because the condition involves a delayed initial retrieval attempt.

Method

Subjects. Forty-eight Washington University undergraduates, ages 18–22 years, participated in Experiment 1 in exchange for course credit.

Design and materials. Experiment 1 used a 2 × 6 mixed factorial design. Retention interval before the final test (10 min vs. 2 days) was manipulated between-subjects, and five different spacing conditions, plus one control condition, were manipulated within-subjects. In three conditions, subjects studied a vocabulary word pair and then took three subsequent tests over that pair. For example, subjects studied *sobriquet–nickname* or *benison–blessing* and later were tested with *sobriquet* and *benison* to recall *nickname* and *blessing*, respectively. In the massed condition, subjects studied a word pair and took three consecutive tests without any other trials intervening between the tests (denoted 0–0–0). In the expanding condition, one trial occurred between the study trial and the first test, five trials occurred between the first and second tests, and nine trials occurred between the second and third tests (1–5–9). In the equally spaced condition, five trials occurred between the study trial and subsequent test trials (5–5–5). Two single-test conditions were also investigated, in addition to the repeated test conditions. In the immediate-test condition, one trial occurred between the study trial and a single test trial, and in the delayed-test condition, five trials occurred between the study and test trial. Finally, the vocabulary word pairs were rotated through a nonstudied control condition, in which the word pair was not studied during the learning phase but was tested on the final test, to estimate subjects' prior knowledge of the vocabulary words.

Fifty-two vocabulary word pairs were selected from test preparation books for the GRE (see Pashler, Zarow, & Triplett, 2003). Each pair consisted of a vocabulary word and a one-word definition (e.g., *sobriquet–nickname*). Thirty-six of the pairs were used as experimental pairs and were divided into six sets of six pairs for counterbalancing, and the other 16 pairs were used only as filler items during the learning phase. A list of 112 trials was created for the learning phase, in which the five spacing conditions occurred six times, with the constraint that the study trial for all five conditions occurred before the next cycle of the five conditions began. This constraint ensured that the conditions were evenly distributed throughout the learning phase. The mean serial positions of the study trials for each condition were roughly equal: massed (0–0–0) = 56.7; expanding (1–5–9) = 49.7; equal (5–5–5) = 50.7; single-immediate (1) = 57.0; single-delayed (5) = 53.0. There were three primacy and three recency buffers in the list of trials. The six sets of vocabulary word pairs were rotated through the six conditions (the five spacing conditions plus the nonstudied control condition), creating six different counterbalancing orders. Four subjects were assigned to each counterbalancing order in the 10-min final test condition, and likewise, 4 subjects were assigned to each counterbalancing order in the 2-day final test condition.

Procedure. Subjects were told that the learning phase consisted of a series of study and test trials distributed throughout the phase. During study trials, subjects saw a vocabulary word with its one-word definition printed below it on a computer screen and were told to study the pair so that they could remember it later on. Each study or test trial lasted 8 s, with a 500-ms intertrial interval. During test trials, subjects saw a vocabulary word with a cursor below it, and their job was to type in the correct definition word for the vocabulary word. Each test trial lasted 8 s, with a 500-ms intertrial interval, and after each test trial the computer program automatically advanced to the next trial, regardless of whether the subject had entered a response. Response latencies were assessed as the duration between the onset of the cue and the last keystroke of the response (which is the default measure of response times for string responses in E-Prime; see Schneider, Eschman, & Zuccolotto, 2002). Subjects were not given feedback about the accuracy of their responses on test trials. After the learning phase, the subjects were engaged in a distracter task (playing a video game on the computer) for 10 min.

Twenty-four of the subjects took the final retention test immediately following the 10-min distracter phase, and the other 24 subjects were dismissed and returned for the final test 2 days after the first session. On the final test, subjects were told that they would see a vocabulary word with a cursor below it, and their job was to type in the correct definition word for each vocabulary word. Subjects were given 14 s for each test trial (with a 500-ms intertrial interval). After completing the final test, the subjects were debriefed and thanked for their participation.

Results

Learning phase recall. An initial analysis of the learning phase results showed no differences in performance between the 10-min and 2-day groups, so the analyses performed on the learning phase recall and response time data were collapsed across retention interval conditions. The top portion of Table 1 shows the mean proportion of items correctly recalled on each test during the learning phase. Not surprisingly, recall in the massed condition was nearly perfect on the three massed tests. Recall was greater on the first test in the expanding retrieval condition, which occurred after only one intervening trial, than on the first test in the equally spaced condition, which occurred after five intervening trials. The single-test conditions showed a similar pattern of results: Recall was greater in the immediate-test condition (1) than in the delayed-test condition (5). The Test 1 recall scores were submitted to a 2 (Spacing of the first test: 1 vs. 5) × 2 (Test Condition: Single vs. Repeated) analysis of variance (ANOVA). There was a main effect of spacing the first test, $F(1, 47) = 5.99$, $\eta_p^2 = .11$, but no effect of test condition and no Spacing × Test Condition interaction (Fs < 1). Increasing the spacing of the first test from one intervening trial to five intervening trials reduced recall on that test (79% vs. 73%; $d = 0.36$; $p_{\text{rep}} = .95$). (p_{rep} is an estimate of the probability of replicating the direction of an effect, described by Killeen, 2005.)

Table 1
Mean Recall and Mean Response Times (in Milliseconds) During the Learning Phase in Experiment 1

	Learning phase recall		
Learning condition	Test 1	Test 2	Test 3
Massed (0–0–0)	.98 (.01)	.98 (.01)	.98 (.01)
Expanding (1–5–9)	.78 (.03)	.76 (.03)	.77 (.03)
Equal (5–5–5)	.73 (.03)	.73 (.03)	.73 (.03)
Single-immediate (1)	.81 (.03)		
Single-delayed (5)	.73 (.03)		

	Learning phase response times		
Learning condition	Test 1	Test 2	Test 3
Massed (0–0–0)	2,592 (64)	2,408 (64)	2,333 (60)
Expanding (1–5–9)	3,428 (112)	3,233 (110)	2,966 (108)
Equal (5–5–5)	3,579 (108)	2,988 (99)	2,716 (84)
Single-immediate (1)	3,432 (85)		
Single-delayed (5)	3,553 (108)		

Note. Standard errors are in parentheses.

Recall changed very little across repeated tests in the massed, expanding, and equally spaced conditions. A 3 (Spacing Condition: Massed, Expanding, or Equal) × 3 (Test Number: 1, 2, or 3) ANOVA revealed a main effect of spacing condition, $F(2, 94) = 38.97$, $\eta_p^2 = .45$, but no effect of test number and no interaction (Fs < 1). Collapsed across the three repeated tests, recall in the massed condition was greater than recall in the expanding condition (98% vs. 77%; $d = 1.44$, $p_{\text{rep}} = 1.00$) and greater than recall in the equally spaced condition (98% vs. 73%; $d = 1.50$; $p_{\text{rep}} = 1.00$). In addition, recall in the expanding condition was greater than recall in the equally spaced condition (77% vs. 73%; $d = 0.19$; $p_{\text{rep}} = .81$). Together, the two analyses of the learning phase recall data show that expanding retrieval practice produced slightly better learning phase recall than equally spaced practice, that this difference was due to giving an immediate first test in the expanding condition, and that there was little change in recall across repeated tests in either condition.

Learning phase response times. The mean response times for correct recalls on each test during the learning phase were analyzed as an index of retrieval fluency, with longer response times indicating greater retrieval difficulty (Benjamin & Bjork, 1996; Koriat & Ma'ayan, 2005). For each condition, response times that were 2.5 standard deviations above or below the mean were trimmed from the analysis, eliminating 1.9% of the data. In addition, subjects who did not have at least one observation (correct recall) per condition were removed from the analysis, eliminating data from 1 subject.

The mean response times are shown in the bottom portion of Table 1. Overall, response times were fastest in the massed condition. In the immediate single-test and expanding retrieval conditions, delaying the first test by one item resulted in slower response times, relative to the massed condition; likewise, delaying the first test by five items in the delayed single-test and equally spaced conditions led to even slower response times, relative to the other conditions. The Test 1 response latencies were submitted to a 2 (Spacing of the first test: 1 vs. 5) × 2 (Test Condition: Single vs. Repeated) ANOVA, which revealed a main effect of spacing the first test, $F(1, 46) = 3.49$, $\eta_p^2 = .07$, but no main effect of test condition and no interaction (Fs < 1). Collapsed across the single and repeated test conditions, increasing the delay of the first test from one item to five items resulted in a 136 ms slowing in response time (3,432 ms vs. 3,566 ms; $p_{\text{rep}} = .90$), indicating that retrieval was more challenging when the first test occurred after a delay of five items.

In addition, response times grew faster across repeated tests in all three conditions. A 3 (Spacing Condition: Massed, Expanding, or Equal) × 3 (Test Number: 1, 2, or 3) ANOVA revealed a main effect of spacing condition, $F(2, 92) = 54.27$, $\eta_p^2 = .54$; a main effect of test number, $F(2, 92) = 60.20$, $\eta_p^2 = .57$; and a significant Spacing × Test Number interaction, $F(4, 184) = 8.19$, $\eta_p^2 = .15$. Of most importance, response times decreased across repeated tests in the expanding and equally spaced retrieval conditions, indicating that retrieval grew easier across repeated tests that were spaced over trials, not more challenging. The finding is contrary to the idea that gradually expanding the interval between repeated tests would make retrieval increasingly difficult (Landauer & Bjork, 1978).

Another interesting effect of delaying the first test is evident in Table 1. When the first test was delayed, response times in the second test were faster than response times on the second test following an immediate first test. In the equally spaced condition, response times on the second test were 248 ms faster than response times on the second test in the expanding retrieval condition (2,988 ms vs. 3,233 ms; $p_{\text{rep}} = .92$). In both conditions, the second test occurred five items after the first test, but when the first test was delayed, the greater retrieval difficulty of the delayed first test led to faster response times on the second test.

Final recall. Of central interest are the effects of repeated testing and spacing schedule (massed, expanding, or equal) on long-term retention. Table 2 shows the mean proportion of items recalled on the final criterial test as a function of spacing condition. Baseline performance in the nonstudied control condition was very low ($M = 1.1\%$; 3 subjects produced the correct definition word for one of the vocabulary words in the control condition), so the data in Table 2 reflect learning during the experiment and not prior knowledge. Several key results are evident in Table 2. First, massed repeated testing produced the worst retention at both retention intervals. Indeed, at both retention intervals, performance in the massed practice condition was slightly worse than performance for both the single-test conditions (with spacings of one or five items before the test). More important, expanding retrieval produced a modest benefit over equally spaced retrieval on the final test given after 10 min, conceptually replicating Landauer and Bjork (1978). However, this result was reversed on the 2-day retention test: Equally spaced retrieval practice led to better delayed retention than expanding retrieval. A 2 (Spacing Condition: Expanding vs. Equal) × 2 (Retention Interval: 10 min vs. 2 days) ANOVA did not reveal a main effect of spacing condition ($F < 1$), but there was a main effect of retention interval, $F(1, 46) = 15.95$, $\eta_p^2 = .26$, and a significant Spacing × Retention Interval interaction, $F(1, 46) = 7.23$, $\eta_p^2 = .14$. Although expanding retrieval practice led to slightly better performance on the 10-min test (71% vs. 62%; $d = 0.30$; $p_{\text{rep}} = .86$), the opposite pattern occurred 2 days later: Equally spaced retrieval practice produced better long-term retention than expanding retrieval on the 2-day final test (45% vs. 33%; $d = 0.50$; $p_{\text{rep}} = .93$).

Table 2
Final Recall in Experiment 1

	Retention interval	
Learning condition	10 min	2 days
Massed (0–0–0)	.47 (.06)	.20 (.04)
Expanding (1–5–9)	.71 (.05)	.33 (.05)
Equal (5–5–5)	.62 (.07)	.45 (.05)
Single-immediate (1)	.65 (.05)	.22 (.04)
Single-delayed (5)	.57 (.06)	.30 (.04)

Note. Standard errors are in parentheses.

A similar pattern of results was observed for the single-test conditions: On the 10-min criterial test, the immediate-test condition (1) produced better retention than the delayed-test condition (5), but on the 2-day criterial test, the delayed-test condition produced better long-term retention. A 2 (Spacing Condition: 1 vs. 5) × 2 (Retention Interval: 10 min vs. 2 days) ANOVA did not reveal a main effect of spacing condition ($F < 1$), but there was a main effect of retention interval, $F(1, 46) = 37.35$, $\eta_p^2 = .45$, and a significant Spacing × Retention Interval interaction, $F(1, 46) = 5.24$, $\eta_p^2 = .10$. The immediate-test condition produced better retention than the delayed-test condition on the 10-min test (65% vs. 57%; $d = 0.29$; $p_{\text{rep}} = .88$), but the delayed-test condition produced better long-term retention on the 2-day final test (30% vs. 22%; $d = 0.41$; $p_{\text{rep}} = .85$).

At both retention intervals, repeated testing in the expanding and equally spaced conditions led to better final recall, relative to taking a single-test. The repeated test data, collapsed across the 1–5–9 and 5–5–5 conditions, and the single-test data, collapsed across the one and five conditions, were submitted to a 2 (Test Condition: Single vs. Repeated) × 2 (Retention Interval: 10 min vs. 2 days) ANOVA. There was a main effect of repeated testing, $F(1, 46) = 11.40$, $\eta_p^2 = .20$; a main effect of retention interval, $F(1, 46) = 30.00$, $\eta_p^2 = .40$; and a significant Test Condition × Retention Interval interaction, $F(1, 46) = 2.11$, $\eta_p^2 = .04$. Repeated testing enhanced final recall at both retention intervals, relative to taking a single test, but the effect of repeated testing was greater at the longer 2-day delay (39% vs. 26%; $d = 0.75$; $p_{\text{rep}} = .99$) than at the shorter 10-min delay (66% vs. 61%; $d = 0.21$; $p_{\text{rep}} = .84$).

Discussion

Although expanding retrieval practice produced a modest benefit in the short term, equally spaced practice led to superior long-term retention 2 days later. Our analyses of performance during the learning phase indicated that having an initial test after a brief delay (a lag of five trials) was responsible for promoting long-term retention relative to testing immediately after study (a lag of one trial). When subjects took just a single test, the delayed-test condition led to better retention than the immediate-test condition, providing converging evidence with the expanding and equally spaced retrieval practice conditions. Delaying the first test made retrieval on that test more difficult, as evidenced by increases in response times. In addition, expanding the interval between repeated tests did not make repeated retrieval more difficult, contrary to the idea that gradually increasing the spacing of tests would increase retrieval difficulty on the tests. Instead, response times grew faster across repeated tests, whereas recall performance remained constant. The results of Experiment 1 support our theory that the important difficulty for enhancing learning is delaying an initial retrieval attempt, rather than expanding the spacing of repeated tests.

Experiment 2

In Experiment 2, we investigated the effects of feedback on expanding and equally spaced retrieval practice at short and long retention intervals. Landauer and Bjork's (1978) theory holds that expanding retrieval is effective because the first test in an expanding schedule of retrieval practice occurs relatively soon after studying. Expanding retrieval promotes higher levels of retrieval success than conditions in which the first test occurs after a delay (such as in equally spaced practice) during which forgetting can occur. However, providing feedback after test trials would counteract forgetting on the delayed first test by allowing subjects to correct errors on repeated tests. Any benefits of increasing the difficulty of retrieval on the delayed first test, however, should still occur even if feedback is provided after the test. Providing feedback also represents what students are most likely to do if they use a spaced retrieval technique to test themselves while they are studying. Experiment 2 was carried out simultaneously with Experiment 1 and was identical to it in all respects but one: After each test trial during the learning phase, subjects were told whether they were correct or incorrect and were shown the correct response.

Method

Subjects. Forty-eight Washington University undergraduates, ages 18–22 years, participated in Experiment 2 in exchange for course credit. None of the subjects had participated in Experiment 1.

Design, materials, and procedure. The design and materials were identical to those used in Experiment 1. The list of study and test trials used in the learning phase of Experiment 2 was the same as the one used in Experiment 1. The procedure was nearly identical, except that subjects were given feedback about their responses after test trials. After each test trial, the word "correct" or "incorrect" was shown on the computer screen, and subjects were shown the vocabulary word and its correct definition for 4 s on the computer screen. As in Experiment 1, half of the subjects took the final criterial test following the 10-min distracter task, and half returned for the final criterial test 2 days later. Although subjects were given feedback during the learning phase, they were not given feedback about any of their responses on the final test, consistent with the procedure used in Experiment 1.

Results

Learning phase recall. As in Experiment 1, an initial analysis of the learning phase results showed no differences in performance between the 10-min and 2-day groups, so the analyses performed on the learning phase recall and response time data were collapsed across retention interval conditions. The top portion of Table 3 shows the mean proportion of items correctly recalled during the learning phase. Recall in the massed condition was near ceiling on all three recall tests. As in Experiment 1, when the first test occurred relatively immediately after studying, in the immediate-test and expanding retrieval conditions, recall was greater than when the first test occurred after a delay, in the delayed-test and equally spaced conditions. The Test 1 recall scores were submitted to a 2 (Spacing of the first test: 1 vs. 5) × 2 (Test Condition: Single vs. Repeated Test) ANOVA. There was a main effect of spacing the first test, $F(1, 47) = 3.56$, $\eta_p^2 = .07$, but no effect of test condition and no Spacing × Test Condition interaction (Fs < 1). Thus, recall performance on the first test in Experiment 2 was nearly identical to performance in Experiment 1, which was expected because feedback was given after the first test and should not have affected performance on the that test. Increasing the spacing of the first test from one to five items reduced recall on the first test (77% vs. 72%; $d = 0.26$; $p_{\text{rep}} = .91$).

Table 3
Mean Recall and Mean Response Times (in Milliseconds) During the Learning Phase in Experiment 2

	Learning phase recall		
Learning condition	Test 1	Test 2	Test 3
Massed (0–0–0)	.98 (.01)	.99 (.01)	.99 (.01)
Expanding (1–5–9)	.77 (.03)	.87 (.03)	.93 (.02)
Equal (5–5–5)	.71 (.03)	.89 (.02)	.94 (.02)
Single-immediate (1)	.77 (.03)		
Single-delayed (5)	.73 (.03)		
	Learning phase response times		
Learning condition	Test 1	Test 2	Test 3
Massed (0–0–0)	2,716 (62)	2,594 (60)	2,540 (68)
Expanding (1–5–9)	3,415 (108)	3,201 (90)	2,940 (84)
Equal (5–5–5)	3,547 (117)	3,036 (89)	2,834 (72)
Single-immediate (1)	3,339 (100)		
Single-delayed (5)	3,568 (104)		

Note. Standard errors are in parentheses.

Providing feedback in Experiment 2 led to increased recall across repeated tests in the expanding and equally spaced conditions but not in the massed condition, in which performance was at ceiling on all three tests. The repeated test scores in the expanding and equally spaced conditions were submitted to a 2 (Spacing Condition: Expanding vs. Equal) × 3 (Test Number: 1, 2, or 3) ANOVA. There was no main effect of spacing condition ($F < 1$), but there was a main effect of test number, $F(2, 94) = 50.28$, $\eta_p^2 = .51$, and a Condition × Test Number interaction, $F(2, 94) = 3.01$, $\eta_p^2 = .06$. When feedback was given during the learning phase in Experiment 2, recall increased across repeated tests, and, despite differences in recall on the first test, performance in the two spaced testing conditions converged near ceiling by the third test.

Learning phase response times. The bottom portion of Table 3 shows the mean response times for correct recalls on each test during the learning phase. For each condition, response times that were 2.5 standard deviations above or below the mean were trimmed from the analysis, eliminating 2.1% of the data. In addition, one subject who did not have at least one observation per condition was removed from further analyses. Overall, response times were fastest in the massed condition (as they were in Experiment 1). Delaying the first test by one item led to slower response times, relative to the massed condition, and delaying the first test by five items produced even slower response times, relative to the other conditions. The Test 1 response latencies were submitted to a 2 (Spacing of the first test: 1 vs. 5) × 2 (Test Condition: Single vs. Repeated) ANOVA. There was a main effect of spacing the first test, $F(1, 46) = 4.51$, $\eta_p^2 = .09$, but no main effect of test condition and no interaction ($F\text{s} < 1$). Collapsed across the single and repeated test conditions, increasing the delay of the first test from one item to five items resulted in a 181-ms slowing in response time (3,377 ms vs. 3,558 ms; $p_{rep} = .93$), indicating greater retrieval difficulty when the first test was delayed.

Response times grew faster across repeated tests in the three repeated test conditions. The response time data were submitted to a 3 (Spacing Condition: Massed, Expanding, or Equal) × 3 (Test Number: 1, 2, or 3) ANOVA. (Data from 1 additional subject, who was missing data in one condition, were removed from this analysis.) There was a main effect of spacing condition, $F(2, 90) = 42.70$, $\eta_p^2 = .49$; a main effect of test number, $F(2, 90) = 48.58$, $\eta_p^2 = .52$; and a significant Spacing × Test Number interaction, $F(4, 180) = 7.43$, $\eta_p^2 = .14$. Providing feedback in Experiment 2 did not change the overall pattern of response times across repeated tests, relative to the pattern of results obtained in Experiment 1 (without feedback). Response times grew faster across repeated tests in the massed, expanding, and equally spaced conditions, indicating that retrieval grew easier across repeated tests, not more difficult. The effect of delaying the first test on response times on the second test, observed in Experiment 1, was again obtained in Experiment 2. When the first test was delayed, response times on the second test were faster than response times on the second test following an immediate first test. In the equally spaced condition, response times on the second test were 165 ms faster than response times on the second test in the expanding retrieval condition (3,036 ms vs. 3,201 ms; $p_{rep} = .91$). Greater retrieval difficulty on the delayed first test led to faster response times on the second test.

Final recall. Table 4 shows the mean proportion of items recalled on the final retention test. Baseline performance in the nonstudied control condition was very low ($M = 0.7\%$; 2 subjects produced the correct definition word for one of the vocabulary words in the control condition). As in Experiment 1, massed repeated testing produced the worst retention at both retention intervals, even slightly worse than both single-test conditions. Expanding retrieval practice led to slightly better performance than equally spaced practice on the 10-min criterial test, although performance in both conditions was near ceiling. However, equally spaced retrieval practice led to better performance than expanding retrieval on the 2-day criterial test, replicating the primary result of Experiment 1. A 2 (Spacing Condition: Expanding vs. Equal) × 2 (Retention Interval: 10 min vs. 2 days) ANOVA revealed a main effect of spacing condition, $F(1, 46) = 2.79$, $\eta_p^2 = .06$; a main effect of retention interval, $F(1, 46) = 37.75$, $\eta_p^2 = .45$; and a significant Spacing × Retention Interval interaction, $F(1, 46) = 6.18$, $\eta_p^2 = .12$. Expanding retrieval practice led to a slight advantage over equally spaced practice on the 10-min test, though ceiling effects cloud interpretation of this result (90% vs. 87%; $d = 0.24$; $p_{rep} = .82$). However, equally spaced practice led to better performance on the 2-day final test (60% vs. 49%; $d = 0.41$; $p_{rep} = .93$), replicating the results of Experiment 1 and other studies that provided feedback during the learning phase and measured retention after a delay (e.g., Cull, 2000).

Table 4
Final Recall in Experiment 2

	Retention interval	
Learning condition	10 min	2 days
Massed (0–0–0)	.49 (.05)	.19 (.05)
Expanding (1–5–9)	.90 (.02)	.49 (.06)
Equal (5–5–5)	.87 (.03)	.60 (.05)
Single-immediate (1)	.60 (.05)	.24 (.04)
Single-delayed (5)	.71 (.05)	.36 (.06)

Note. Standard errors are in parentheses.

A 2 (Spacing Condition: 1 vs. 5) × 2 (Retention Interval: 10 min vs. 2 days) ANOVA was performed on the single-test data. There was a main effect of spacing condition, $F(1, 46) = 14.13$, $\eta_p^2 = .24$, and a main effect of retention interval, $F(1, 46) = 29.53$, $\eta_p^2 = .39$, but no Spacing × Retention Interval interaction ($F < 1$). In Experiment 2, when feedback was provided after the tests, the delayed-test condition outperformed the immediate-test condition at both the 10-min (71% vs. 60%; $d = 0.44$; $p_{rep} = .95$) and 2-day (36% vs. 24%; $d = 0.48$; $p_{rep} = .97$) retention intervals.

A final analysis investigated the effects of repeated testing in the expanding and equally spaced conditions, relative to taking a single test. The repeated test data, collapsed across the 1–5–9 and 5–5–5 conditions, and the single-test data, collapsed across the one and five conditions, were submitted to a 2 (Test Condition: Single vs. Repeated) × 2 (Retention Interval: 10 min vs. 2 days) ANOVA. There was a main effect of repeated testing, $F(1, 46) = 78.93$, $\eta_p^2 = .63$, and a main effect of retention interval, $F(1, 46) = 40.83$, $\eta_p^2 = .47$. No Test Condition × Interval Interaction was observed ($F < 1$). The analysis confirmed that repeated testing in the expanding and equally spaced conditions led to better retention than taking a single test at both retention intervals.

Discussion

In Experiment 2, expanding retrieval led to a slight benefit on the 10-min final test, although performance was near ceiling in both conditions. However, equally spaced practice led to better long-term retention after a 2-day delay. The single-test conditions provided converging evidence to show that delaying the first test led to better retention, and when feedback was given to counteract forgetting on the test, this result was observed at both retention intervals. Performance during the learning phase again indicated that delaying the first test promoted long-term retention. When the first test occurred after a brief delay, retrieval on the test was more difficult, as evidenced by increases in response times. Furthermore, as we observed in Experiment 1, expanding the interval between repeated tests did not make repeated retrieval more difficult, but instead response times grew faster across repeated tests. The results of Experiment 2 provide converging evidence with Experiment 1 and confirm our prediction that delaying an initial retrieval attempt promotes long-term retention, especially when feedback is given after the test to counteract forgetting that occurs on the delayed first test.

Experiment 3

Experiments 1 and 2 showed that equally spaced retrieval practice led to better performance than expanding retrieval on a delayed criterial test, a surprising result inconsistent with the common belief that expanding retrieval promotes long-term retention (Bjork, 1988; Landauer & Bjork, 1978) but consistent with our theory that delaying an initial retrieval attempt represents the important difficulty for promoting long-term retention. The implications of Experiments 1 and 2 are that delaying the first retrieval attempt enhances long-term retention and that the particular schedule of repeated tests does not matter much. However, when expanding and equally spaced schedules of retrieval practice are compared, the position of the first test is naturally confounded with the schedule of repeated tests. Expanding schedules of retrieval practice involve a first test relatively immediately after studying (a spacing of one item in Experiments 1 and 2), whereas equally spaced schedules involve a first test after a brief delay (a spacing of five items in Experiments 1 and 2). The purpose of Experiment 3 was to separate the effects of delaying the first test and the schedule of repeated tests to examine whether expanding the schedule of repeated tests is the key factor for enhancing long-term retention (Landauer & Bjork, 1978) or whether delaying the first test is the key, regardless of how repeated tests are spaced, as our theory predicts.

One prior study by Carpenter and DeLosh (2005) also noted that the position of the first test and the schedule of repeated tests are confounded in the standard comparison of expanding and equally spaced conditions. In one of their experiments, Carpenter and DeLosh compared a 3–3–3 condition with a 3–5–7 condition and thereby equated the position of the first test but varied the spacing of the second and third initial tests. They found no difference between the two conditions on a final recall test immediately after the learning phase. Their result is consistent with our hypothesis that the delay before the first retrieval attempt is crucial for determining retention, whereas the particular schedule of repeated tests is less critical. In Experiment 3 in the present research, we examined our theory using more conditions than those used by Carpenter and DeLosh and assessed performance on a final criterial test at two retention intervals (10 min or 2 days).

In Experiment 3, four repeated test conditions were used to separate the effects of spacing the first test (immediate vs. delayed) and the effect of the schedule of repeated tests (expanding vs. equally spaced). In two immediate test conditions, the first test occurred after a spacing of 0 items, to maximize retrieval success on this test. One immediate-test condition then involved three repeated tests spaced according to an expanding schedule (0–1–5–9), and the other immediate-test condition involved three equally spaced repeated tests (0–5–5–5). In two delayed test conditions, the first test occurred after a delay of five items. One delayed-test condition then involved three expanding tests (5–1–5–9), and the other delayed-test condition involved three equally spaced tests (5–5–5–5). We also examined two single-test conditions (a 0 condition and a 5 condition), similar to the single-test conditions in the previous experiments. No feedback was provided after the test trials. As in the previous experiments, half of the subjects took a criterial test after 10 min, and half took a criterial test 2 days later.

Method

Subjects. Fifty-six Washington University undergraduates, ages 18–22 years, participated in Experiment 3 in exchange for course credit. None of the subjects had participated in the previous two experiments.

Design and materials. A 2 × 7 mixed factorial design was used in Experiment 3. Retention interval before the final test (10 min vs. 2 days) was manipulated between-subjects, and six different spacing conditions, plus one control condition, were manipulated within-subjects. Four of the spacing conditions involved repeated testing, in which subjects studied a vocabulary word pair and then took four tests over the pair. Two of the repeated test conditions involved an immediate first test, which occurred after an initial spacing of zero items. In one immediate-test condition, the three repeated tests (Tests 2–4) were spaced according to an expanding schedule: The second test occurred one trial after the first test, the third test occurred five trials after the second test, and the fourth test occurred nine trials after the third test (0–1–5–9). In the other immediate-test condition, Tests 2–4 were equally spaced: Five trials occurred between each of the repeated tests (0–5–5–5). The other two repeated test conditions involved a delayed first test, which occurred after an initial spacing of five items. One delayed-test condition involved an expanding schedule of repeated tests (5–1–5–9), and the other delayed-test condition involved equally spaced repeated tests (5–5–5–5). Two single-test conditions were also investigated to provide converging evidence with the repeated test conditions about the effect of delaying the first test. One single-test condition involved an immediate test, after a spacing of zero items, and the other single-test condition involved a delayed test, after a spacing of five items. The vocabulary word pairs were rotated through a nonstudied control condition, in which the word pair was not studied during the learning phase but was tested on the final test.

Fifty-six vocabulary word pairs were used in Experiment 3. Four vocabulary words were added to the set of 52 words used in Experiments 1 and 2. Forty-two of the pairs were used as experimental pairs and were divided into seven sets of six pairs for counterbalancing, and the other 14 pairs were used only as filler items during the learning phase. For the learning phase, a list of 161 trials was created in which the six spacing conditions occurred six times, with the constraint that the study trial for all six conditions occurred before the next cycle of the six conditions began. The mean serial positions of the study trial for each condition were as follows: immediate-expanding (0–1–5–9) = 69.2; immediate-equal (0–5–5–5) = 67.0; delayed-expanding (5–1–5–9) = 70.2; delayed-equal (5–5–5–5) = 69.0; immediate-single (0) = 69.5; delayed-single (5) = 73.2. There were two primacy and two recency buffers in the list of trials used in Experiment 3. The seven sets of vocabulary word pairs were rotated through the seven conditions (the six spacing conditions plus the nonstudied control condition), creating seven different counterbalancing orders. Four subjects were assigned to each counterbalancing order in the 10-min final test condition, and likewise, 4 subjects were assigned to each counterbalancing order in the 2-day final test condition.

Procedure. The procedure was identical to the one used in Experiment 1, in which no feedback was given after test trials. Subjects were told that the learning phase consisted of a series of study and test trials distributed throughout the learning period.

During study trials, subjects saw a vocabulary word with its one-word definition printed below it on a computer screen and were told to study the pair so that they could remember it later on. During test trials, subjects saw a vocabulary word with a cursor below it, and their job was to type in the correct definition word for each vocabulary word. Subjects were not given feedback about the accuracy of their responses after test trials. After the learning phase, the subjects were engaged in a distracter task (playing a video game on the computer) for 10 min. Twenty-eight of the subjects took the final test immediately after the 10-min distracter phase, and the other 28 subjects took the final test 2 days later. The final test was identical to the test trials in the learning phase, in which subjects were given the vocabulary words and were asked to recall the definition words.

Results

Learning phase recall. An initial analysis of the learning phase results showed no differences in performance between the 10-min and 2-day groups, so the learning phase recall and response time results were collapsed across retention interval condition. Table 5 shows the mean proportion of items correctly recalled on each test during the learning phase. Recall was greater on the first test when the test occurred immediately than when it occurred after a delay, replicating the results of Experiments 1 and 2. The Test 1 data were submitted to a 2 (Spacing of the first test: 0 vs. 5) × 3 (Condition: Single-Test, Expanding, or Equal) ANOVA. There was a main effect of spacing the first test, $F(1, 55) = 113.92$, $\eta_p^2 = .67$, but no main effect of test condition and no interaction (Fs < 1). Increasing the delay of the first test from zero items to five items reduced recall on that test (97% vs. 72%; $d = 1.80$; $p_{rep} = 1.00$).

A second analysis investigated the effect of delaying the first test on recall on the repeated tests (Tests 2, 3, and 4). An initial analysis of the repeated test data confirmed that there were no significant changes in recall across Tests 2–4 ($F < 1$), so the data were collapsed across test number. The results were submitted to a 2 (Spacing of the first test: 0 vs. 5) × 2 (Spacing of repeated tests: Expanding vs. Equal) ANOVA. There was a main effect of delaying the first test, $F(1, 55) = 3.19$, $\eta_p^2 = .06$, and a marginal effect of spacing condition, $F(1, 55) = 2.44$, $\eta_p^2 = .04$, but no interaction ($F < 1$). Overall, having an immediate first test led to better recall on Tests 2, 3, and 4 than having a delayed first test (78% vs. 74%; $d = 0.21$; $p_{rep} = .89$). In addition, performance was slightly greater on Tests 2–4 when the tests were spaced according to an expanding schedule, rather than equally spaced (77% vs. 74%; $d = 0.16$; $p_{rep} = .86$).

Learning phase response times. Table 5 also shows the mean response times for correct recalls on each test during the learning phase. For each condition, response times that were 2.5 standard deviations above or below the mean were trimmed from the analysis, eliminating 1.8% of the data. Two subjects did not have at least one observation (correct recall) per condition and were removed from further analyses. Response times were faster on the first test when the test occurred immediately after studying, relative to when the first test occurred after a delay. The Test 1 response latencies were submitted to a 2 (Spacing of the first test: 0 vs. 5) × 3 (Test Condition: Expanding, Equal, or Single) ANOVA. There was a main effect of spacing the first test, *F*(1,

Table 5
Mean Recall and Mean Response Times (in Milliseconds) During the Learning Phase in Experiment 3

	Learning phase recall			
Learning condition	Test 1	Test 2	Test 3	Test 4
Immediate initial test				
Expanding (0–1–5–9)	.97 (.01)	.81 (.02)	.79 (.03)	.78 (.03)
Equal (0–5–5–5)	.98 (.01)	.74 (.03)	.77 (.03)	.77 (.03)
Delayed initial test				
Expanding (5–1–5–9)	.73 (.03)	.74 (.03)	.76 (.03)	.75 (.03)
Equal (5–5–5–5)	.71 (.03)	.73 (.03)	.72 (.03)	.72 (.03)
Single test conditions				
Immediate (0)	.96 (.01)			
Delayed (5)	.72 (.03)			

	Learning phase response times			
Learning condition	Test 1	Test 2	Test 3	Test 4
Immediate initial test				
Expanding (0–1–5–9)	2,459 (57)	2,839 (81)	2,712 (67)	2,729 (71)
Equal (0–5–5–5)	2,476 (53)	2,932 (77)	2,728 (76)	2,543 (64)
Delayed initial test				
Expanding (5–1–5–9)	3,257 (105)	2,732 (85)	2,615 (67)	2,576 (87)
Equal (5–5–5–5)	3,260 (71)	2,721 (65)	2,540 (65)	2,547 (58)
Single test conditions				
Immediate (0)	2,630 (54)			
Delayed (5)	3,224 (100)			

Note. Standard errors are in parentheses.

53) = 141.79, η_p^2 = .73, but no main effect of test condition ($F <$ 1) and no interaction, F(2, 106) = 1.55. Collapsed across the single, expanding, and equally spaced test conditions, delaying the first test led to a 725-ms slowing in response times relative to immediate testing (2,521 ms vs. 3,247 ms; p_{rep} = 1.00), indicating that greater retrieval effort was required when the first test occurred after a delay of five items.

Response times also grew faster across repeated tests (Tests 2, 3, and 4), replicating the pattern of results observed in Experiments 1 and 2. The results were submitted to a 2 (Spacing of the first test: 0 vs. 5) × 2 (Test Condition: Expanding vs. Equal) × 3 (Test Number: 2, 3, or 4) ANOVA. (One additional subject, who was missing data in one of the conditions, was removed from this analysis.) There was a main effect of delaying the first test, F(1, 52) = 8.41, η_p^2 = .14, and a main effect of test number, F(2, 104) = 13.34, η_p^2 = .20, but no effect of spacing condition ($F <$ 1). In addition, none of the interactions reached significance (Fs $<$ 1). Overall, response times grew faster across repeated tests, regardless of whether the tests were expanding or equally spaced. In addition, delaying the first test facilitated response times on the repeated tests, relative to when the first test occurred immediately. Collapsed across spacing conditions (expanding vs. equally spaced) and collapsed across repeated tests (Tests 2–4), response times on the repeated tests were 125 ms faster when the first test was delayed than when it occurred immediately (2,622 ms vs. 2,749 ms; p_{rep} = .98).

Final recall. Table 6 shows the mean proportion of items recalled on the final retention test. Baseline performance in the nonstudied control condition was very low (M = 1.2%; 4 subjects produced the correct definition word for one of the vocabulary words in the control condition). The differences among the four repeated test conditions on the 10-min final test were small. A 2 (Spacing of the first test: 0 vs. 5) × 2 (Repeated Testing: Expanding vs. Equal) ANOVA performed on the 10-min final test data did not reveal a main effect of spacing the first test ($F <$ 1), or a main effect of the schedule of repeated tests ($F <$ 1), or a significant interaction, F(1, 27) = 1.18. Collapsed across the expanding and equally spaced conditions, there was a slight advantage of the immediate-test conditions over the delayed-test conditions, in the same direction as the results of Experiment 1 (64% vs. 62%; d = 0.08; p_{rep} = .67). However, on the 2-day final test, the conditions with a delayed first test (5–1–5–9 and 5–5–5–5) led to better performance than the conditions with an immediate initial test (0–1–5–9 and 0–5–5–5). A 2 (Spacing of the first test: 0 vs. 5) × 2 (Repeated Testing: Expanding vs. Equal) ANOVA performed on the 2-day final test data revealed a main effect of spacing the first test, F(1, 27) = 6.26, η_p^2 = .19, but no effect of the schedule of repeated tests and no significant interaction (Fs $<$ 1). Collapsed across the expanding and equally spaced conditions, the conditions with a delayed first test led to better long-term retention than the conditions with an immediate first test (52% vs. 44%; d = 0.35; p_{rep} = .94), regardless of whether repeated tests were expanding or equally spaced.

Table 6
Final Recall in Experiment 3

Learning Condition	Retention interval 10 min	2 days
Immediate initial test		
Expanding (0–1–5–9)	.62 (.05)	.43 (.06)
Equal (0–5–5–5)	.66 (.05)	.45 (.06)
Delayed initial test		
Expanding (5–1–5–9)	.63 (.05)	.51 (.05)
Equal (5–5–5–5)	.61 (.05)	.52 (.05)
Single test conditions		
Immediate (0)	.36 (.05)	.21 (.04)
Delayed (5)	.52 (.06)	.39 (.06)

Note. Standard errors are in parentheses.

The data from the single-test conditions were submitted to a 2 (Spacing of the first test: 0 vs. 5) × 2 (Retention Interval: 10 min vs. 2 days) ANOVA. There was a main effect of spacing the first test, F(1, 54) = 25.59, η_p^2 = .32, and a main effect of retention interval, F(1, 54) = 4.78, η_p^2 = .08, but no Spacing × Retention Interval interaction ($F <$ 1). At both retention intervals, the delayed-test conditions led to better retention than the immediate-test conditions.

A final analysis investigated the effects of repeated testing, collapsed across the expanding and equally spaced conditions, relative to taking a single test. Separate analyses were performed on the 10-min and 2-day data. The 10-min results were submitted to a 2 (Spacing of the first test: 0 vs. 5) × 2 (Test Condition: Single vs. Repeated) ANOVA. There was a main effect of spacing the first test, F(1, 27) = 6.59, η_p^2 = .20; a main effect of repeated testing, F(1, 27) = 29.63, η_p^2 = .52; and a significant spacing of the First Test × Repeated Test interaction, F(1, 27) = 8.63, η_p^2 = .24. Overall, the repeated test conditions led to better performance than taking a single test, and the interaction was driven by the difference between the immediate and delayed single-test conditions favoring the former condition. A similar analysis was performed on the 2-day final test results. A 2 (Spacing of the first test: 0 vs. 5) × 2 (Test Condition: Single vs. Repeated) ANOVA revealed a main effect of spacing the first test, F(1, 27) = 21.45, η_p^2 = .44, with a delayed first test leading to better performance than an immediate initial test, a main effect of repeated testing, F(1, 27) = 28.11, η_p^2 = .51, but a nonsignificant spacing of the First Test × Repeated Test interaction, F(1, 27) = 2.19, η_p^2 = .08. Together, the analyses confirm that repeated testing enhanced retention relative to taking a single test and that delaying the first test was critical to improved performance.

Discussion

Experiment 3 separated the effects of delaying the first test from the effects of the schedule of repeated tests, which are confounded in typical comparisons of expanding versus equally spaced retrieval practice conditions. When these two factors were disentangled, the results showed that delaying the first test enhanced long-term retention on the 2-day final test, regardless of the schedule of repeated tests. Performance on the tests during the learning phase was similar to performance in the previous experiments. Delaying the first test by five trials increased the difficulty of retrieval on that test, as evidenced by longer response latencies on the delayed test. As we observed in the previous experiments, expanding the interval between repeated tests did not make repeated retrieval increasingly more difficult, as indexed by response latency. Experiment 3 clearly shows that delaying an initial retrieval attempt represents the important difficulty for enhancing learning, rather than expanding the schedule of repeated tests.

General Discussion

The present experiments support our hypothesis that delaying an initial retrieval attempt (as is done in equally spaced retrieval practice conditions) promotes long-term retention by increasing the difficulty of retrieval on the first test. In Experiments 1 and 2, we found a modest short-term benefit of expanding retrieval practice over equally spaced retrieval on a retention test 10 min after learning, replicating previous studies that used similar procedures and short retention intervals (Cull et al., 1996; Landauer & Bjork, 1978; but see Balota et al., 2006; Carpenter & DeLosh, 2005). However, on a final test 2 days after learning, equally spaced retrieval produced better long-term retention than expanding retrieval (Cull, 2000). This pattern of results occurred with or without feedback during the initial learning phase (Experiments 2 and 1, respectively). Experiment 3 showed that the positive effect of equally spaced practice was due to having the first test after a delay. When the spacing of the first test and the spacing of repeated tests were factorially crossed, the results showed that delaying the first test enhanced long-term retention, regardless of whether repeated tests were expanding or equally spaced. Although these results contradict the conventional wisdom that expanding retrieval represents a superior form of retrieval practice (Landauer & Bjork, 1978), they are consistent with the hypothesis that delaying the first test represents a desirable difficulty for promoting learning.

We discuss our results first in relation to previous research on expanding and equally spaced retrieval practice. We then discuss implications of our work for theories of the testing effect, especially the idea that retrieval difficulty promotes long-term retention. Finally, we discuss the implications of our results for memory training procedures and for educational practices.

Relation to Prior Work on Expanding and Equally Spaced Retrieval Practice

The present experiments examined two factors that previous research suggested might mediate the effects of expanding and equally spaced retrieval practice: Whether feedback was provided after tests, and whether the final criterial test occurred after a short retention interval (within the same session) or after a long one (in a separate, delayed session). Regarding feedback during the learning phase, Cull et al. (1996, Experiment 5) gave subjects feedback and found no difference between expanding and equally spaced practice on an immediate criterial test, given a few seconds after the learning phase. Cull (2000) also gave subjects feedback and found a slight benefit of equally spaced practice over expanding retrieval on criterial tests immediately after learning, and he also observed greater benefits of equally spaced practice on criterial tests given 3 days and 8 days later. In the present experiments, we also found no differences between expanding and equally spaced practice on an immediate criterial test when feedback was given during learning, although performance was near ceiling in both conditions. On 2-day delayed criterial tests, however, equally spaced retrieval practice led to better performance than expanding retrieval when feedback was given, although we also observed this result in Experiment 1 without feedback. Providing feedback after each test counteracted forgetting that occurred in the equally spaced condition when the first test was delayed, allowing subjects to correct errors and improve performance on repeated tests. Interestingly, providing feedback did not appear to affect the difficulty of retrieval on each test, as indicated by the response time results. Thus, the positive effects of delaying the first test in the equally spaced condition remained intact, and feedback supplied the added benefit of allowing subjects to correct errors on repeated tests.

Another effect of feedback was evident in our data looking across Experiments 1 and 2. Feedback enhanced retention in the expanding and equally spaced conditions (cf. Table 2 to Table 4) at both retention intervals, but in contrast feedback did not make massed practice any more effective than it was without feedback. Performance in the massed condition was virtually identical in the two experiments on the 10-min test (47% vs. 49%) and on the 2-day test (20% vs. 19%). In other words, the positive effects of feedback observed in the other conditions were neutralized when repeated tests were massed. Seven massed study and test trials (counting feedback trials in the feedback condition) were no more beneficial than four massed trials (in the no feedback condition). In fact, the data in Table 4 show that seven massed trials were worse for long-term retention than a single delayed test followed by feedback (19% vs. 36%; $d = 0.66$; $p_{\mathrm{rep}} = .98$).

The second variable that mediates the effects of spaced retrieval, the retention interval before the final test, is more surprising in light of the widely held belief that expanding retrieval promotes long-term retention. Of the previous studies comparing expanding and equally spaced practice, only Cull (2000) investigated effects on a delayed final test in two of his experiments, and in both experiments he found positive effects of equally spaced retrieval over expanding retrieval. We obtained the same result in Experiments 1 and 2: Equally spaced practice led to better performance than expanding retrieval on a delayed criterial test. In Experiment 3, we separated the effects of spacing the first test and the schedule of repeated tests, which are confounded in typical comparisons of expanding and equally spaced retrieval practice. The present results go beyond prior research and show that delaying the first test is responsible for the positive effects of equally spaced practice on long-term retention, regardless of the schedule of repeated tests.

In our experiments, we used a rather limited range of possible spacing schedules, although this criticism could be levied against previous research as well. Logan and Balota (in press) recently compared several equally spaced and expanding schedules of retrieval practice (for example, 1–2–3 vs. 2–2–2; 1–3–5 vs. 3–3–3; 1–3–8 vs. 4–4–4) and also tested both younger and older adults. Importantly, Logan and Balota gave subjects a final criterial test either immediately after the learning phase or after a 24-hr delay. They found that expanding retrieval led to better performance during the learning phase than equally spaced retrieval practice, but this initial benefit was lost on a final test given at the end of the learning phase (although older adults showed a slight benefit from expanding retrieval). However, on the final test given 24 hr later, equally spaced retrieval produced better long-term retention than expanding retrieval for both younger and older adults. Our results agree with Logan and Balota's showing that equally spaced retrieval practice leads to better long-term retention than expanding retrieval.

Most prior research has not examined differences between expanding and equally spaced retrieval practice by analyzing performance during the learning phase. Landauer and Bjork's (1978) theory is that expanding retrieval practice promotes retention because retrieval success is high on an immediate first test, and expanding the interval between repeated tests increases the retrieval difficulty involved in those tests. In the present experiments, although success was high when the first test occurred immediately after studying (spacing of zero or one trials), we found no evidence that gradually expanding the spacing of repeated tests made retrieval more difficult on those tests. Instead, response times grew faster across repeated tests in the expanding and equally spaced conditions, indicating that retrieval grew increasingly easy across repeated tests. Logan and Balota (in press) also examined response latencies during the learning phase and obtained similar results: Response times grew faster across repeated tests in the expanding and equally spaced practice conditions. Furthermore, when the position of the first test was equated and the repeated tests were expanding or equal in Experiment 3 (e.g., comparing 0–1–5–9 with 0–5–5–5) we found no differences in response times across the expanding and equal tests (Tests 2–4). This pattern of results contradicts the assumption that retrieval becomes increasingly more difficult across repeated tests under expanding retrieval practice conditions.

In sum, the evidence is mixed regarding the effect of expanding versus equally spaced retrieval practice at short retention intervals, with approximately half of the existing experiments showing an advantage of expanding over equal spacing (e.g., Cull et al., 1996; Landauer & Bjork, 1978; our present results) and half showing no effect or an opposite advantage of equally spaced practice (Balota et al., 2006; Carpenter & DeLosh, 2005; Cull, 2000). In contrast, all previous experiments examining long-term retention on a delayed criterial test have showed benefits of equally spaced retrieval practice over expanding retrieval (Cull, 2000; Logan & Balota, in press; our present results). We know of no existing study using a continuous paired associate learning task (following Landauer & Bjork, 1978) that has shown that expanding retrieval produces greater long-term retention (after delays greater than 24 hr) than equally spaced practice.

Retrieval Difficulty and the Testing Effect

The finding that increasing retrieval difficulty on an initial test promotes learning is consistent with other prior research showing that increasing processing difficulty improves later retention (for reviews, see Bjork, 1999; McDaniel & Einstein, 2005; Roediger & Karpicke, 2006a). For example, prior research has shown that difficult retrieval is positively correlated with retention on a later criterial test. Gardiner et al. (1973) had subjects answer general knowledge questions and measured the amount of time it took them to answer each question. When subjects took longer to answer the questions, indicating that retrieving the answer was difficult, final free recall of the answers was greater for the more difficult questions than for questions that took shorter times to answer, indicating easier retrieval (see too Benjamin et al., 1998).

The study by Gardiner et al. (1973) relied on the inherent difficulty of questions in determining retrieval difficulty. Other research has experimentally manipulated the degree of retrieval difficulty on an initial test and found positive effects. For example, Whitten and Bjork (1977) varied the amount of time subjects spent performing a Brown–Peterson distracter task between when they studied a word pair and when they attempted recall on a first test. They found that when subjects spent longer amounts of time performing the distracter task before initial retrieval, thereby increasing retrieval difficulty, final recall of the word pairs was greater than when subjects spent a shorter amount of time performing the distracter task (see also Bjork & Allen, 1970; Modigliani, 1976). Jacoby (1978) manipulated the difficulty of a first test by having either 0 or 20 intervening items occur between a study trial and a test trial, using a continuous paired associate task. He ensured relatively high retrieval success on the test by giving subjects a fragment of the target word as an additional retrieval cue on the test. On a final criterial test, Jacoby showed a large benefit of spacing the first test and argued that effortful processing on the delayed test was responsible for the effect.

One idea about why an immediate first test does not promote long-term retention is that when a test trial occurs shortly after a study trial, the item is recalled from primary memory, and prior research has shown that recall from primary memory does not translate into long-term retention. For example, when subjects study a list of items and then recall them in any order on an immediate test, subjects recall items at the end of the list better than items from the middle of the list, and this recency effect in single-trial free recall is thought to reflect recall from primary memory. However, if subjects are given a final free recall test after a delay (e.g., at the end of the experimental session), performance shows a negative recency effect, and items from the end of the list are recalled worse than items from the middle of the list (see Bjork, 1975; Craik, 1970; Madigan & McCabe, 1971). The negative recency effect shows that recalling items from primary memory does not enhance long-term retention.

Another idea about why an immediate first test does not enhance retention has been discussed by Balota et al. (2006, 2007). They noted that when the first test occurs immediately in an expanding schedule of retrieval practice, it is essentially a massed repetition, and the immediate test may be ineffective for promoting long-term retention for the same reasons that massed repetition leads to poor retention. Our experiments clearly showed that massed retrieval practice leads to poor long-term retention, even worse than a single delayed test followed by feedback (in Experiment 2). Balota et al. have explained the positive effect of equally spaced retrieval practice in terms of encoding variability (Bower, 1972; Estes, 1955; Martin, 1968; see too Crowder, 1976, chapter 9). Briefly, the encoding variability explanation assumes that performance on a memory test depends on the overlap between contextual elements at study and at test. When a first test occurs immediately after study, there has been little time for contextual elements to fluctuate between the study trial and the test trial, and performance will be high when short-term retention is assessed on final test relatively immediately after learning. However, when a first test occurs after a delay, contextual variability will increase because contextual elements will fluctuate between study and test. When long-term retention is assessed after a delay, presumably after contextual elements have fluctuated to an even greater extent, performance will be best in the delayed-test condition because the likelihood that contextual elements during learning will overlap with contextual elements on the final test will be greatest in this condition. Thus, delaying a first test enhances contextual variability on the test, thereby promoting long-term retention (see Balota et al., 2007).

Other explanations of the testing effect have emphasized the role of retrieval processes in producing the effect (see Roediger & Karpicke, 2006a). A broad explanation of the testing effect is based on the idea that tests enhance learning when they require greater depth of retrieval, an idea similar to the notion of depth of processing at encoding (Craik & Tulving, 1975). According to this account, tests enhance long-term retention when they promote effortful retrieval. Evidence for the effortful retrieval explanation of the testing effect comes from studies that have investigated the effects of different test formats (recall vs. recognition) on later retention. The results of several experiments show that the testing effect is greater when an initial test involves recall, or production of information, than when it involves recognition, or identification, presumably because recall involves more effortful retrieval than recognition (see Butler & Roediger, in press; Carpenter & DeLosh, 2006; Glover, 1989; Kang, McDermott, & Roediger, in press; McDaniel, Anderson, Derbish, & Morrisette, in press). Delaying an initial test also increases the effortful retrieval involved on the test. Our analyses of performance during the learning phase showed that response times were slower when the first test occurred after a brief delay, indicating that delaying the test made retrieval more effortful. Thus, just as recall tests produce bigger testing effects by promoting effortful retrieval, delaying an initial retrieval attempt may also enhance retention by effortful retrieval.

Implications for Memory Training and for Enhancing Student Learning

Expanding retrieval has been advocated as a memory enhancement technique for older adults and memory-impaired populations, such as individuals with DAT (Camp, 2006). The technique is thought to be effective for these populations because it ensures retrieval success on an initial test and gradually shapes production of the desired response at increasingly longer intervals, with the goal of promoting long-term retention. Camp and his colleagues (Camp, 2006; Camp et al., 2000) have suggested that expanding retrieval might be effective because it promotes errorless learning, a method aimed at reducing or eliminating the production of errors during learning that was originally developed to train animals in discrimination learning tasks (Deutsch & Terrace, 1967; Karpicke & Hearst, 1975; Terrace, 1963, 1966). Wilson, Baddeley, Evans, and Sheil (1994) have argued that errorless learning is especially important in training programs for individuals with memory impairments. Our results have only indirect implications for such programs, but they suggest that conditions that produce the fewest errors during learning (like massed practice in Experiments 1 and 2) can lead to very poor long-term retention. Delaying an initial test, thereby making an initial retrieval attempt more difficult, leads to more errors on the test but also produces the biggest gains in long-term retention, especially when feedback is given to correct errors committed on the delayed first test (see too Pashler et al., 2003). We think that comparison of specific schedules of spaced practice may be important for determining which memory training techniques are most effective, and on the basis of our results, equally spaced practice appears to be more effective than expanding retrieval. Given the implications of spaced retrieval for memory training practices, it is surprising that there have not been more systematic investigations of different schedules of spaced retrieval practice (see Balota et al., 2007).

The present results are perhaps more directly relevant to student learning and educational practices, because we tested college students using educationally relevant materials in a task similar to what students might do when they are studying (in this case, to improve their vocabulary before taking the GRE). Some authors have argued that students should implement expanding retrieval practice to enhance their learning (Cull et al., 1996; Rea & Modigliani, 1985), but our results clearly show that equally spaced retrieval leads to better long-term retention than expanding retrieval when conditions typical of prior research are compared (Experiments 1 and 2). Furthermore, when the position of the first test and the spacing of repeated tests were factorially crossed in Experiment 3, delaying the first retrieval attempt enhanced retention regardless of whether repeated tests were expanding or equally spaced. Instructors may wish to advocate expanding retrieval practice because it often promotes recall success during learning because of the initial retrieval attempt soon after studying. However, our results indicate that this feature of expanding retrieval, a relatively immediate first test, is exactly what renders the technique less effective than equally spaced practice. Furthermore, if students self-test on their own, they will likely give themselves feedback after each test, and feedback will counteract any forgetting that would occur when the first retrieval attempt is delayed and unsuccessful.

In addition to the fact that expanding retrieval produced inferior long-term retention relative to equally spaced retrieval practice, one other drawback to expanding retrieval is worth noting: We found that constructing sequences of trials that conformed to expanding spacing schedules was a tedious and often frustrating task. It may be much easier for students simply to space their initial retrieval attempt, thereby making initial retrieval more difficult, and then repeatedly test themselves, though the particular schedule of repeated tests does not matter much.

We hasten to add, however, that delaying a first test may not benefit learning when the delay is too long. When a first test occurs a long time after learning, and performance on the test is poor, delayed testing clearly would not confer as much benefit as testing more immediately after learning. Delaying a first test also may not benefit learning of more complex prose materials, as opposed to paired associate materials used in the present experiments and nearly all prior research on expanding retrieval. For example, in a classic study of the testing effect, Spitzer (1939) had students read prose passages and then gave them multiple-choice tests after varying delays. He found that testing immediately after studying prevented forgetting on the test and promoted performance on later repeated tests. In another experiment using prose materials, we also found that taking a free recall test relatively immediately after studying a passage promoted long-term retention more than having the first test occur after a short 10-min delay (Karpicke & Roediger, 2006). Presumably, when subjects recall material from a prose passage, little or none of their performance reflects recall from primary memory; thus, the test may be inherently difficult enough to facilitate learning and need not be delayed to make it more difficult. However, in our study and in Spitzer's, students were not given feedback after the tests. If students receive feedback after tests, as they would if they were testing themselves and spacing their practice over several days in preparation for an upcoming exam, then a delayed first test might benefit learning more than an immediate test, despite poorer performance on the delayed test (see too Pashler et al., 2003). The issue of what the optimal delay should be before a first retrieval attempt to promote learning of different types of materials awaits future research.

Conclusion

Practicing retrieval is a powerful way to enhance learning and retention. However, the present results indicate that expanding retrieval, a popular method of implementing retrieval practice, may not be the most effective spacing technique for improving long-term retention. Instead, equally spaced practice leads to better long-term retention because the condition involves a first test after a brief delay, and the greater effort involved in the initial test enhances later retention. We believe our results are in line with Bjork's (1994, 1999) concept of creating desirable difficulties to enhance learning. However, equally spaced retrieval practice creates a desirable difficulty that enhances learning more than expanding retrieval practice. The important difficulty for promoting long-term retention is a delayed initial retrieval attempt, not expanding the interval between repeated tests.

References

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

Balota, D. A., Duchek, J. M., & Logan, J. M. (2007). Is expanded retrieval practice a superior form of spaced retrieval? A critical review of the extant literature. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger, III* (pp. 83–106). New York: Psychology Press.

Balota, D. A., Duchek, J. M., & Paullin, R. (1989). Age-related differences in the impact of spacing, lag, and retention interval. *Psychology and Aging, 4,* 3–9.

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

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. Reder (Ed.), *Implicit memory and metacognition* (pp. 309–338). Hillsdale, NJ: Erlbaum.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127,* 55–68.

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Bjork, R. A. (1988). Retrieval practice and the maintenance of knowledge. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 1, pp. 396–401). New York: Wiley.

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

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII. Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Bjork, R. A., & Allen, T. W. (1970). The spacing effect: Consolidation or differential encoding? *Journal of Verbal Learning and Verbal Behavior, 9,* 567–572.

Bower, G. H. (1972). Stimulus sampling theory of encoding variability. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory* (pp. 85–123). New York: Wiley.

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

Camp, C. J. (2006). Spaced retrieval: A model for dissemination of a cognitive intervention for persons with dementia. In D. K. Attix & K. A. Welsh-Bohmer (Eds.), *Geriatric neuropsychology: Assessment and intervention* (pp. 275–292). New York: Guilford Press.

Camp, C. J., Bird, M. J., & Cherry, K. E. (2000). Retrieval strategies as a rehabilitation aid for cognitive loss in pathological aging. In R. D. Hill, L. Backman, & A. S. Neely (Eds.), *Cognitive rehabilitation in old age* (pp. 224–248). New York: Oxford University Press.

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology, 19,* 619–636.

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

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

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

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval induced facilitation: Initially nontested material can benefit from prior testing. *Journal of Experimental Psychology: General, 135,* 553–571.

Craik, F. I. M. (1970). The fate of primary memory items in free recall. *Journal of Verbal Learning and Verbal Behavior, 9,* 143–148.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104,* 268–294.

Craik, F. I. M., & Watkins, M. J. (1973). The role of rehearsal in short-term memory. *Journal of Verbal Learning and Verbal Behavior, 12,* 599–607.

Crowder, R. G. (1976). *Principles of learning and memory.* Hillsdale, NJ: Erlbaum.

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

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

Deutsch, J. A., & Terrace, H. S. (1967, May 19). Discrimination learning and inhibition. *Science, 156,* 988–989.

Estes, W. K. (1955). Statistical theory of distributional phenomena in learning. *Psychological Review, 62,* 369–377.

Gardiner, J. M., Craik, F. I. M., & Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition, 1,* 213–216.

Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior, 15,* 1–16.

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

Hochhalter, A. K., Overmier, J. B., Gasper, S. M., Bakke, B. L., & Holub, R. J. (2005). A comparison of spaced retrieval to other schedules of practice for people with dementia. *Experimental Aging Research, 31,* 101–118.

Hogan, R. M., & 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.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17,* 649–667.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (in press). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology.*

Karpicke, J., & Hearst, E. (1975). Inhibitory control and errorless discrimination learning. *Journal of the Experimental Analysis of Behavior, 23,* 159–166.

Karpicke, J. D., & Roediger, H. L. (2006). *Does expanding retrieval work with prose materials?* Unpublished manuscript, Washington University in St. Louis.

Karpicke, J. D., & Roediger, H. L. (in press). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language.*

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32,* 1–24.

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science, 16,* 345–353.

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52,* 478–492.

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

Logan, J. M., & Balota, D. A. (in press). Expanded vs. equal interval spaced retrieval practice: Exploration of schedule of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition.*

Madigan, S. A. (1969). Intraserial repetition and coding processes in free recall. *Journal of Verbal Learning and Verbal Behavior, 8,* 828–835.

Madigan, S. A., & McCabe, L. (1971). Perfect recall and total forgetting: A problem for models of short-term memory. *Journal of Verbal Learning and Verbal Behavior, 10,* 101–106.

Martin, E. (1968). Stimulus meaningfulness and paired-associate transfer: An encoding variability hypothesis. *Psychological Review, 75,* 421–441.

Matvey, G., Dunlosky, J., & Guttentag, R. (2001). Fluency of retrieval at study affects judgments of learning (JOLs): An analytic or nonanalytic basis for JOLs? *Memory & Cognition, 29,* 222–232.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (in press). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology.*

McDaniel, M. A., & Einstein, G. O. (2005). Material appropriate difficulty: A framework for determining when difficulty is desirable for improving learning. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications* (pp. 73–85). Washington, DC: American Psychological Association.

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 371–385.

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.

Modigliani, V. (1976). Effects on a later recall by delaying initial recall. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 609–622.

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.

Peterson, L. R., Wampler, R., Kirkpatrick, M., & Saltzman, D. (1963). Effect of spacing of presentations on retention of paired-associates over short intervals. *Journal of Experimental Psychology, 66,* 206–209.

Rea, C. P., & Modigliani, V. (1985). The effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. *Human Learning, 4,* 11–18.

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210.

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

Schacter, D. L., Rich, S. A., & Stampp, M. S. (1985). Remediation of memory disorders: Experimental evaluation of the spaced retrieval technique. *Journal of Clinical and Experimental Neuropsychology, 7,* 79–96.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3,* 207–217.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime reference guide.* Pittsburgh, PA: Psychology Software Tools.

Serra, M. J., & Dunlosky, J. (2005). Does retrieval fluency contribute to the underconfidence-with-practice effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1258–1266.

Shaughnessy, J. J., & Zechmeister, E. B. (1992). Memory-monitoring accuracy as influenced by the distribution of retrieval practice. *Bulletin of the Psychonomic Society, 30,* 125–128.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30,* 641–656.

Starch, D. (1927). *Educational psychology.* Oxford, England: Macmillan.

Terrace, H. S. (1963). Discrimination learning with and without "errors." *Journal of the Experimental Analysis of Behavior, 6,* 1–27.

Terrace, H. S. (1966, December 30). Discrimination learning and inhibition. *Science, 154,* 1677–1680.

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.

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. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3,* 240–245.

Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior, 16,* 465–478.

Wilson, B. A., Baddeley, A. D., Evans, J., & Sheil, A. (1994). Errorless learning in the rehabilitation of memory impaired people. *Neuropsychological Rehabilitation, 4,* 307–326.

Received November 28, 2006
Revision received March 7, 2007
Accepted March 12, 2007 ■