Is expanding retrieval a superior method for learning text materials?

JEFFREY D. KARPICKE
Purdue University, West Lafayette, Indiana

AND

HENRY L. ROEDIGER III
Washington University, St. Louis, Missouri

Expanding retrieval practice refers to the idea that gradually increasing the spacing interval between repeated tests ought to promote optimal long-term retention. Belief in the superiority of this technique is widespread, but empirical support is scarce. In addition, virtually all research on expanding retrieval has examined the learning of word pairs in paired-associate tasks. We report two experiments in which we examined the learning of text materials with expanding and equally spaced retrieval practice schedules. Subjects studied brief texts and recalled them in an initial learning phase. We manipulated the spacing of the repeated recall tests and examined final recall 1 week later. Overall we found that (1) repeated testing enhanced retention more than did taking a single test, (2) testing with feedback (restudying the passages) produced better retention than testing without feedback, but most importantly (3) there were no differences between expanding and equally spaced schedules of retrieval practice. Repeated retrieval enhanced long-term retention, but how the repeated tests were spaced did not matter.

We know that practicing retrieval enhances long-term retention (e.g., Karpicke & Roediger, 2008; Roediger & Karpicke, 2006a), and we also know that spacing repetitions enhances retention (e.g., Glenberg, 1976; Greene, 1989, 2008; Rohrer & Pashler, 2007). Combining the benefits of spacing and retrieval practice ought to produce especially powerful effects on learning. This article is concerned with identifying the best way to space retrieval practice to promote learning of educationally relevant text materials.

A straightforward idea is simply to maximize the spacing interval between a study episode and subsequent repeated tests. It is true that increasing the spacing intervals between study periods tends to produce increasing gains in learning, especially when retention is assessed at long intervals (Glenberg, 1976; Greene, 2008). But a problem occurs when retrieval attempts are spaced over time: Increasing the spacing of tests also increases the likelihood of forgetting on the tests. If subjects have forgotten material by the time they attempt retrieval, the attempt would likely not produce a positive effect on later retention unless the subjects were given feedback and another retrieval opportunity.

Landauer and Bjork (1978) offered expanding retrieval practice as a remedy to this problem (see too Spitzer, 1939). The idea behind expanding retrieval is that people should retrieve information immediately after studying it and then gradually increase the interval between repeated retrieval attempts. Landauer and Bjork examined the learning of face–name pairs in a continuous paired-associate task and manipulated the number of trials that occurred between each study and test trial. In an expanding condition the subjects studied a face–name pair and then recalled it immediately (with 0 intervening trials), recalled it again after 1 more trial, then again after 3 more trials and again after 8 more trials. This expanding retrieval condition was denoted 0-1-3-8 to indicate the number of trials that occurred between each study and test trial. Landauer and Bjork compared the expanding condition to an equally spaced condition (3-3-3-3) wherein the repeated tests were evenly spaced and the total spacing was equivalent to that of the expanding condition (a total spacing of 12 trials). Both spaced retrieval conditions produced better performance on a final recall test 30 min after the learning phase relative to when the subjects studied the pair only once and did not practice retrieval. But the important result was that the expanding condition produced about a 10% advantage over the equally spaced condition on the final test.

The explanation for the superiority of expanding retrieval is that the technique combines the benefits of retrieval success and retrieval difficulty. Recalling information on an immediate first test ensures retrieval success on the test, and gradually expanding the interval between repeated recall attempts is intended to increase the difficulty of those attempts. Retrieval difficulty itself is assumed to produce a positive effect on long-term retention (see Gardiner, Craik, & Bleasdale, 1973; Pyc & Rawson, 2009).

J. D. Karpicke, karpicke@purdue.edu

The idea of expanding retrieval is intuitive, and belief in the efficacy of the method caught on soon after the publication of Landauer and Bjork's (1978) chapter. But the empirical evidence endorsing the superiority of expanding retrieval is mixed. Most researchers have examined the effect of expanding retrieval on retention after relatively short retention intervals, following the lead of Landauer and Bjork. But the results of replication attempts have been equivocal. For instance, Cull, Shaughnessy, and Zechmeister (1996) did not find significant effects of expanding retrieval over equally spaced retrieval in three of five experiments with short retention intervals (see too Carpenter & DeLosh, 2005; Pyc & Rawson, 2007). In addition, experiments that used longer delays (e.g., 24 h or more) often found no effect of expanding retrieval, or sometimes an advantage of equally spaced practice (e.g., Cull, 2000; Logan & Balota, 2008). And finally, some experiments in which subjects were given feedback after retrieval attempts also found no advantage of expanding retrieval or an advantage of equally spaced practice (for a review, see Balota, Duchek, & Logan, 2007).

We recently carried out a series of experiments to examine the effects of expanding and equally spaced retrieval schedules at short and long retention intervals (Karpicke & Roediger, 2007a). We had students learn vocabulary word pairs and manipulated whether repeated tests were massed (a 0-0-0 schedule), expanding (a 1-5-9 schedule), or equally spaced (a 5-5-5 schedule). At a short delay (10 min after learning) we replicated Landauer and Bjork's (1978) finding of an advantage of expanding retrieval. However, 2 days after the learning phase, the equally spaced retrieval condition produced better retention than did expanding retrieval. We also examined performance with feedback after each test and again found that equally spaced practice was superior to expanding retrieval at the 2-day interval. In short, our paired-associate learning experiments led us to conclude that expanding retrieval was not superior to equally spaced retrieval and, in fact, the opposite may be true for long-term retention.

Why would equally spaced retrieval practice produce better long-term retention than expanding retrieval? Our theory was that the critical factor in paired-associate learning experiments was the position of the first retrieval attempt, not the schedule of repeated tests. In expanding retrieval conditions the first retrieval attempt occurs almost immediately after studying an item (lags of 0 or 1 trial). We reasoned that this retrieval attempt is often not effective because retrieval occurs while the encoded cue and target items still reside in immediate memory. There is considerable evidence that retrieval of items in the focus of immediate awareness does not support later retention (e.g., Jacoby, 1978). Therefore, equally spaced retrieval practice might enhance retention because that schedule involves a delayed first test (e.g., a lag of 5 trials between study and a first test). However, the position of the first retrieval attempt and the schedule of repeated retrieval attempts are confounded in typical comparisons of expanding and equally spaced conditions. Expanding retrieval conditions involve an immediate first test (e.g., 1-5-9) and equally spaced conditions involve a delayed first test (e.g., 5-5-5). We therefore carried out a final experiment—again with paired-associate materials—in which we factorially crossed the position of the first test (immediate or delayed) and the spacing of repeated tests (expanding or equal). We found that delaying the first test enhanced long-term retention but the schedule of repeated tests did not matter (Karpicke & Roediger, 2007a).

In the present article we investigated the effects of different schedules of retrieval practice on the learning and retention of educationally relevant expository texts. There have been a few recent experiments examining text learning with spaced rereading (Rawson & Kintsch, 2005) and with retrieval practice (Chan, McDermott, & Roediger, 2006; Roediger & Karpicke, 2006b) but there have been no published studies examining how the particular spacing of retrieval attempts affects text learning. In two experiments subjects read brief text passages and practiced reconstructing their knowledge of the texts on free recall tests. We manipulated the spacing of the recall tests. In both experiments we factorially crossed the position of the first test (immediate or delayed) and the spacing of repeated tests (expanding or equal). We examined the effects of the different retrieval practice schedules on a final criterial test 1 week after the learning phase. Our goal was to determine whether there were any meaningful differences between expanding and equally spaced retrieval schedules for the learning of text materials.

The conclusion from our previous research was that delaying a first retrieval attempt produced better long-term retention than did using an immediate first test because, on immediate tests, recall reflected immediate or primary memory (Karpicke & Roediger, 2007a). However, important differences exist between recalling individual items in paired-associate tasks and recalling expository texts in free recall tasks. It is clear that recall reflects immediate/primary memory when a word pair is recalled immediately after it was studied. Similarly, when subjects freely recall lists of items like pictures or words, they often recall the last few items first (see Deese & Kaufman, 1957; Tulving, 1968) and these items are not well retained on delayed tests (cf. the negative recency effect; Craik, 1970). However, in our previous studies of text recall (Roediger & Karpicke, 2006b) subjects did not adopt this strategy when recalling expository texts. That is, the subjects did not tend to recall the last few ideas in the text first, but instead typically recalled ideas in forward serial order, perhaps striving to maintain the narrative or structural coherence of the text. Therefore when students recall ideas from texts, a test might nominally occur immediately after study, but functionally the first ideas recalled are not the last few ideas residing in immediate memory. To examine subjects' recall strategies we analyzed the order in which the subjects recalled ideas from the texts. We expected our analyses to confirm that the subjects recalled the texts in serial order rather than recalling the last few ideas first. If true, this observation would confirm that immediate recall operates quite differently in the recall of text material than in the recall of paired associates or word lists.

EXPERIMENT 1

In Experiment 1, subjects read brief text passages and took three initial free recall tests. We manipulated how the recall tests were spaced in the initial learning session. Specifically, we factorially crossed (1) whether the first test occurred immediately after reading the passage or after a brief delay filled with other activities and (2) whether the repeated tests were expanding or equally spaced. This manipulation created four conditions: immediate-expanding, immediate-equal, delayed-expanding, and delayed-equal. Prior research on expanding and equally spaced retrieval has typically compared two of the four conditions: the immediate-expanding and delayed-equal conditions. We examined performance in all four conditions and assessed long-term retention on a final free recall test 1 week after the learning phase. If expanding retrieval is a superior form of spaced retrieval practice, then the two expanding conditions (immediate-expanding and delayed-expanding) should produce better long-term retention than the two equally spaced conditions (immediate-equal and delayed-equal). Alternatively, the particular schedule of repeated tests might not affect retention, or equal spacing might even be advantageous, as was found in some previous studies (e.g., Karpicke & Roediger, 2007a; Logan & Balota, 2008).

Method

Subjects. Eighty Washington University undergraduates participated in Experiment 1 in exchange for course credit.

Materials. Two brief text passages were selected from Encyclopedia Britannica (obtained online at www.eb.com). Each passage covered a single topic (*Bamboo* and *Vanuatu*) and was divided into 20 idea units for scoring purposes. The passages were 162 and 155 words in length, respectively.

Design. The initial learning session comprised a series of 4-min periods in which the subjects studied and recalled the texts. There were four learning conditions. Two involved an immediate first test. In the immediate-expanding condition the subjects studied a passage and took the first recall test in the next period (with 0 intervening periods), then a second test after two more periods, and a third test after four more periods (thus a 0-2-4 spacing schedule). (The intervening intervals were filled with study or recall of other passages or with an engaging video game, as specified below.) In the immediate-equal condition the subjects took an immediate first test and then two equally spaced repeated tests (0-2-2). The other two conditions were the same as the first two but involved a delayed first test. In the delayed-expanding condition the subjects took the first test after two intervening periods, then a second after two more periods, and a third after four more (2-2-4). In the delayed-equal condition the subjects took the first test two periods after studying, and then there were two intervening periods between each repeated recall test (2-2-2). The learning condition was manipulated between subjects, and 20 subjects were assigned to each condition.

The learning phase consisted of a series of 4-min periods. In each period the subjects studied a passage, took a recall test, or played a video game filler task. The order in which the subjects studied the two passages was counterbalanced to create two counterbalancing orders within each condition. The schedule of the study and test events for each condition is shown in Appendix A.

Procedure. Subjects were tested in groups of 5 or fewer in two sessions. In Session 1 they were told that they would study two prose passages and take recall tests on them in a series of several 4-min periods. During study periods the subjects studied a passage shown on a computer screen. During test periods, the subjects were told to write down as much as they could remember from a cued passage in

a response book given to them. To signal the beginning of a test period, a "Recall [Passage]" cue appeared on the screen and the screen color changed from white to blue. When 30 sec remained in the test period, the screen color changed again from blue to red and the message "30 seconds left" appeared. At the end of each test period the subjects were told to turn to the next page in their response book, and they were instructed not to look at any previous pages at any time during the experiment. Preliminary pilot testing showed that 4 min was enough time for subjects to reach asymptote in amount recalled (cf. Roediger & Thorpe, 1978).

The periods between the study and test periods for one passage were usually filled with study or test periods for the other passage. But there were some unfilled periods, and during these filler periods the subjects played an engaging video game (*Pac-Man*) on the computer for 4 min.

Session 2 occurred 1 week after Session 1. In Session 2 the subjects took a final recall test over each passage. Each final test lasted 8 min to ensure that the subjects had enough time to express their knowledge on the week-delayed test. At the end of the experiment, the subjects were debriefed and thanked for their participation.

Results

All results, unless otherwise stated, were significant at the .05 level. An initial analysis of the data showed no differences among the passages or counterbalancing orders, and therefore the results have been collapsed across these variables.

Scoring. The free recall data were scored by giving one point for each correctly recalled idea unit (cf. Roediger & Karpicke, 2006b). Initially, 40 recall tests were scored by two raters, and the Pearson product–moment correlation between their scores was $r = .98$. Given the high interrater reliability, the remaining recall tests were scored by one rater.

Initial recall. Table 1 shows the mean proportion of idea units recalled on each test in the initial learning phase. In general the two groups with an immediate first test recalled more idea units than did the two groups with a delayed first test. Recall did not change much across repeated tests, and the particular schedule of tests did not much matter. The data were submitted to a 2 (position of the first test: immediate or delayed) $\times$ 2 (spacing schedule: expanding or equal) $\times$ 3 (test number: 1, 2, or 3) ANOVA. There was a main effect of initial test position [$F(1,76) = 20.27$, $\eta_p^2 = .21$], indicating that delaying the first test produced forgetting. There was also a main effect of test number [$F(2,152) = 4.43$, $\eta_p^2 = .06$], indicating that a modest amount of hypermnesia occurred across the three tests. There was no significant

Table 1
Mean Proportions of Idea Units Recalled During the Learning Phase and on the Final Retention Test (1 Week Later) in Experiment 1

Learning Condition	Learning Phase							
	Test 1		Test 2		Test 3		Final Test	
	M	SE	M	SE	M	SE	M	SE
Immediate First Test								
Expanding (0-2-4)	.75	.02	.75	.02	.75	.02	.66	.03
Equal (0-2-2)	.77	.03	.77	.02	.77	.02	.65	.03
Delayed First Test								
Expanding (2-2-4)	.60	.03	.61	.03	.63	.04	.53	.03
Equal (2-2-2)	.63	.04	.63	.04	.64	.04	.57	.04

main effect of spacing schedule ($F < 1$), and there were no significant interactions ($Fs < 1$).

Final recall. The rightmost column of Table 1 shows final recall. In general the final recall data followed a pattern similar to that of the initial recall data. The subjects who took an immediate initial recall test outperformed those who took a delayed initial test, but the particular spacing of repeated tests did not appear to matter. A 2 (position of the first test) $\times$ 2 (spacing schedule) ANOVA confirmed these observations. There was a main effect of initial test position [$F(1,76) = 10.07$, $\eta_p^2 = .12$] but no effect of spacing schedule and no interaction ($Fs < 1$).

Recall order. Recall order was analyzed to examine the strategies that the subjects used to recall the texts—specifically whether the subjects recalled the last few ideas in the text first or if they instead recalled in serial order. Figure 1 shows the mean recall output position plotted as a function of the position of ideas in the texts. The data have been collapsed across the two texts, the repeated tests, and the spacing conditions. The figure clearly shows that the subjects tended to recall the ideas from the texts in serial order. We also calculated the Pearson correlation between recall output order and input order for every test taken by each subject. The overall average correlation was $r = .89$. The serial recall retrieval strategy did not depend on the text or the test number or the spacing condition. In fact there was not a single subject who recalled the last ideas in the passages first. Therefore it was not the case that the subjects merely recalled the most recent ideas first and then proceeded to recall other ideas from the texts. Instead the subjects tended to recall the ideas in the order in which they occurred in the texts.

Discussion

The pattern of results in Experiment 1 was straightforward. The two groups with immediate first tests recalled more in the learning phase and more on the final test than did the two groups with delayed first tests. It did not matter whether the repeated tests were expanding or equally spaced. At first blush the results may appear inconsistent with our theory that an immediate first test is not mnemonically effective (Karpicke & Roediger, 2007a). However, our analysis of retrieval strategies showed that the subjects recalled ideas from the texts in serial order rather than recalling the most recent items first. Thus it was not the case that the subjects recalled the most recent ideas residing in immediate memory first, as is the case when subjects recall an item immediately after its presentation in a paired-associate learning task.

EXPERIMENT 2

Experiment 2 was designed with three main purposes in mind. First, we sought to replicate Experiment 1 with a design that provides a clearer separation of the position of the first test from the spacing of repeated tests. To achieve this, half of the subjects took an immediate first test with no intervening periods between study and test, and the other half took a delayed first test with two intervening periods. Then the spacing of three repeated tests was either expanding (a 1-2-3 schedule) or equally spaced (a 2-2-2 schedule). This factorially crossed the position of the first test and the spacing of repeated tests to create four conditions: 0-1-2-3, 0-2-2-2, 2-1-2-3, and 2-2-2-2.

Second, we included two control conditions against which we could compare the spaced testing conditions—a study-once condition and a single test condition. This allowed us to examine both the effect of taking a single test relative to studying once and the effect of repeated testing relative to taking just a single test (Karpicke & Roediger, 2007b, 2008). The single test was given either immediately or after two intervening periods.

Third, we examined the effects of feedback. Feedback has potent effects on learning (Butler, Karpicke, & Roediger, 2007, 2008; Pashler, Cepeda, Wixted, & Rohrer, 2005), but we know of only a few studies that examined feedback in the form of rereading a text after students attempted recall (e.g., McDaniel, Howard, & Einstein, 2009). Subjects in the feedback conditions restudied the texts after each initial recall test (and were informed at the beginning of the experiment that they would receive feedback), whereas the subjects in the no-feedback conditions did not. Of course, examining feedback is practically relevant, because if students practice retrieval on their own, they would likely restudy after attempting recall (Karpicke, Butler, & Roediger, 2009). But examining feedback also allows us to determine the generality of any potential advantage of expanding retrieval or lack thereof. Feedback should counteract the forgetting that may occur after a delayed test and permit the subjects to correct errors on subsequent tests. Thus it could be possible that greater retrieval effort on a delayed first test followed by feedback would produce a greater mnemonic effect than an immediate first test with feedback.

We made a number of other minor changes to fit the procedure into an hour-long learning session. First, the
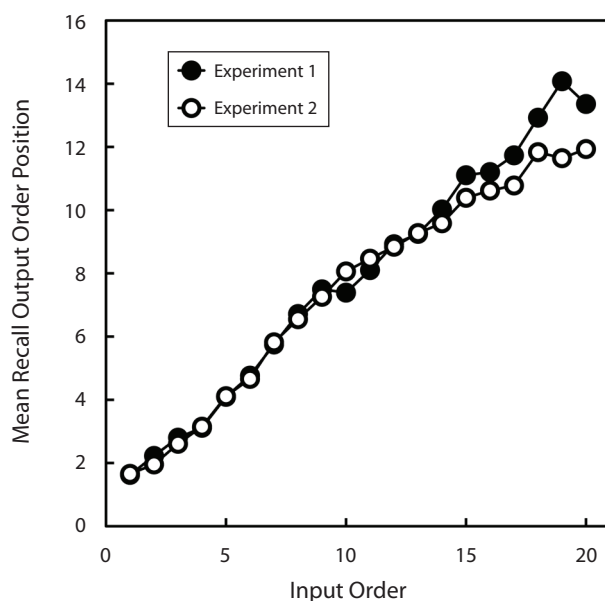

Figure 1. Mean recall output order position as a function of input order (the order in which the 20 ideas occurred in the text passages).

amount of time in each period was reduced to 3 min, and the subjects were instructed to study critical passages just one time and press the space bar when finished. Having subjects study just once was intended to reduce the level of performance in the learning phase to allow us to examine feedback effects without concern for ceiling effects. Second, subjects typed their recall responses on the computer. Finally, subjects read filler passages during the filler periods rather than playing a video game as a distractor task.

Method

Subjects. Sixty-four Washington University undergraduates participated in Experiment 2 in exchange for course credit. None had participated in Experiment 1.

Materials. A total of nine text passages were used in Experiment 2. Four passages were labeled critical passages. These included the two passages used in Experiment 1 (*Bamboo* and *Vanuatu*) and two new passages (*Diving* and *Eugene O'Neill*). The two new passages were selected from a test preparation book (Rogers, 2001; see Roediger & Karpicke, 2006b) and were 170 and 172 words, respectively, and were divided into 20 idea units each for scoring purposes. Five additional passages of approximately the same length as the critical passages were selected from the test preparation book and used as filler passages.

Design. A 4 (learning condition) × 2 (position of the first test: immediate or delayed) × 2 (feedback or no feedback) mixed factorial design was used. There were four learning conditions: study only, single test, expanding retrieval, and equally spaced retrieval. The learning conditions were manipulated within subjects, and one of the four critical passages was assigned to each condition. The four critical passages were presented in the same order for every subject (*Bamboo*, *Diving*, *Eugene O'Neill*, *Vanuatu*), but the order of the four learning conditions was counterbalanced to create four orders. The details of the four counterbalancing orders are shown in Appendix B.

The position of the first test (immediate or delayed) was manipulated between subjects: Half of the subjects took immediate first tests, and the other half took delayed first tests. The presence of feedback was also manipulated between subjects: Half of the subjects restudied the passage for 1 min after each recall test, and half did not. Thus, there were four between-subjects groups, and 16 subjects were assigned to each group.

Procedure. The procedure was similar to the one in Experiment 1. In Session 1, the subjects were told that they would study and recall a series of text passages and that the passages would be shown in either black or blue on the computer screen. They were told that they would be tested on the four critical passages (in black) but not the filler passages (in blue). During study periods for the critical passages, the subjects were told to read the passage one time and press the space bar when they had finished to advance to the next period. Subjects were given a maximum of 3 min to read the passage before the computer automatically advanced. During study periods for the filler passages, the subjects were told to skim the passages for the duration of the 3-min study period. During test periods the subjects saw a "Recall [Passage]" cue on the screen and recalled as much as they could from the passage by typing their responses in a response box shown on the screen. The computer automatically advanced to the next period after 3 min. Preliminary pilot testing showed that this was enough time for subjects to type their responses and reach asymptote in amount recalled. Subjects in the no-feedback group were treated in the same way as the subjects in Experiment 1. Subjects in the feedback group were told that they would study the passage for 1 min after each initial recall test. One week later, in Session 2, the subjects took final free recall tests over each passage. The subjects were given 7 min to recall each passage in the final test phase.

Table 2
Mean Proportions of Idea Units Recalled on Each Test in the Learning Phase in the No-Feedback Conditions in Experiment 2

	Learning Phase							
	Test 1		Test 2		Test 3		Test 4	
Learning Condition	M	SE	M	SE	M	SE	M	SE
Immediate First Test								
Single test (0)	.56	.04						
Expanding (0-1-2-3)	.57	.04	.59	.04	.59	.05	.57	.05
Equal (0-2-2-2)	.58	.05	.53	.05	.57	.05	.54	.06
Delayed First Test								
Single test (2)	.46	.04						
Expanding (2-1-2-3)	.45	.03	.47	.04	.48	.03	.49	.04
Equal (2-2-2-2)	.48	.04	.52	.05	.43	.04	.53	.04

Results

An initial analysis of the data showed no differences among the four passages or the four counterbalancing orders. Therefore the results have been collapsed across passages and orders.

Scoring. As in Experiment 1, 40 recall tests from Experiment 2 were scored by two raters. The Pearson product–moment correlation between their scores was $r = .97$. Given the high interrater reliability, the remaining recall tests were scored by one rater.

Initial recall: No feedback. Table 2 shows the mean proportion of idea units recalled in the no-feedback conditions. The immediate test group recalled more than the delayed test group during the learning phase, and the advantage was due to increased recall on the immediate first test. A 2 (position of the first test) × 3 (learning condition) ANOVA showed that there was a main effect of the position of the first test [$F(1,30) = 4.69$, $\eta_p^2 = .14$] but no main effect of learning condition and no interaction (Fs < 1). The data from the expanding and equally spaced conditions were collapsed across the initial test conditions and entered into a 2 (spacing schedule: expanding or equal) × 4 (test number) ANOVA. There was a marginally significant main effect of test number [$F(3,93) = 2.32$, $\eta_p^2 = .07$, $p = .08$], reflecting a small degree of hypermnesia across tests. But there was no main effect of spacing schedule and no interaction (Fs < 1).

Initial recall: Feedback. Table 3 shows recall in the feedback conditions. The pattern of results was somewhat different from the pattern in the no-feedback conditions.

Table 3
Mean Proportions of Idea Units Recalled on Each Test in the Learning Phase in the Feedback Conditions in Experiment 2

	Learning Phase							
	Test 1		Test 2		Test 3		Test 4	
Learning Condition	M	SE	M	SE	M	SE	M	SE
Immediate First Test								
Single test (0)	.43	.05						
Expanding (0-1-2-3)	.51	.03	.69	.03	.75	.03	.79	.03
Equal (0-2-2-2)	.47	.05	.58	.04	.74	.04	.76	.04
Delayed First Test								
Single test (2)	.43	.04						
Expanding (2-1-2-3)	.47	.04	.70	.04	.80	.03	.81	.04
Equal (2-2-2-2)	.48	.05	.68	.04	.76	.04	.82	.03

First, there did not appear to be a difference in performances on the initial test in the immediate and delayed conditions. A 2 (position of the first test) $\times$ 3 (learning condition) ANOVA confirmed that there was no main effect of initial test on Test 1 recall ($F < 1$), no main effect of condition [$F(2,60) = 1.25$, n.s.], and no interaction ($F < 1$). We can only speculate about why this occurred. Perhaps the subjects knew that they would receive feedback and thus did not expend as much effort recalling on the first test (although this only happened in the immediate test group, which is puzzling).

The data from the expanding and equally spaced conditions were collapsed across the initial test conditions and entered into a 2 (spacing schedule: expanding vs. equal) $\times$ 4 (test number) ANOVA. There was a main effect of test number [$F(3,93) = 154.41$, $\eta_p^2 = .83$], reflecting the fact that recall increased across repeated tests with feedback. There was no main effect of spacing schedule [$F(1,30) = 1.65$, n.s.] and no interaction [$F(3,93) = 1.02$, n.s.].

Final recall. Table 4 shows the final test data with the no-feedback groups in the left column and the feedback groups in the right column. There are several important findings in this table. First, within the no-feedback groups, both the immediate and delayed first test conditions showed significant testing effects. Taking a single test produced better long-term retention than did only studying the text—even when the subjects did not reread the text after recalling it (cf. Roediger & Karpicke, 2006b). A 2 (study only or single test) $\times$ 2 (immediate or delayed test group) ANOVA showed that there was a main effect of testing [$F(1,30) = 17.47$, $\eta_p^2 = .37$]. The main effect of position of the first test was not significant [$F(1,30) = 2.69$, $\eta_p^2 = .08$, $p = .11$], and there was no interaction ($F < 1$). There were also significant testing effects in the feedback groups. A 2 (study only or single test) $\times$ 2 (immediate or delayed test group) ANOVA was performed on the feedback data and showed that there was a main effect of testing [$F(1,30) = 34.51$, $\eta_p^2 = .54$] and a main effect of position of the first test [$F(1,30) = 4.63$, $\eta_p^2 = .13$] but no interaction ($F < 1$).

Second, repeated testing enhanced learning beyond taking just a single test. This occurred in both the feedback and no-feedback conditions. The data from the two repeated test conditions (expanding and equal) were averaged and compared with the single test conditions. The no-feedback data were entered in a 2 (position of the first test) $\times$ 2 (single or repeated test) ANOVA. There was a main effect of repeated testing [$F(1,30) = 38.25$, $\eta_p^2 = .56$] but no main effect of position of the first test [$F(1,30) = 1.05$, n.s.] and no interaction ($F < 1$). The same 2 $\times$ 2 ANOVA was carried out on the feedback data. There was a main effect of repeated testing [$F(1,30) = 264.16$, $\eta_p^2 = .90$] and a main effect of position of the first test [$F(1,30) = 4.72$, $\eta_p^2 = .14$], but no interaction [$F(1,30) = 2.07$, n.s.].

Third—and most important—there were no discernible differences in the effect of expanding and equally spaced retrieval practice on long-term retention. The data from the no-feedback conditions were submitted to a 2 (position of the first test) $\times$ 2 (spacing schedule: expanding or equal) ANOVA. There was no main effect of spacing schedule ($F < 1$), no main effect of position of the first test ($F < 1$), and no interaction [$F(1,30) = 2.22$, n.s.]. The same 2 $\times$ 2 ANOVA was carried out on the feedback data. Again there was no main effect of spacing condition ($F < 1$), no main effect of position of the first test [$F(1,30) = 1.35$, n.s.], and no interaction ($F < 1$). Thus how the repeated tests were spaced did not influence long-term retention.

We carried out a final analysis to compare the no-feedback and feedback conditions. First, the single test data were entered in a 2 (feedback vs. no feedback) $\times$ 2 (position of the first test) ANOVA. There was no significant overall effect of feedback [$F(1,30) = 1.90$, n.s.] and no main effect of position of the first test ($F < 1$), but there was a significant interaction [$F(1,60) = 4.69$, $\eta_p^2 = .07$]. The benefit of feedback in the single test condition appeared only when the first test occurred after a delay. A pairwise analysis showed that there was no effect of feedback for the immediate first test groups ($F < 1$), but there was an effect for the delayed test groups [$F(1,30) = 6.20$, $\eta_p^2 = .17$]. Second, the repeated test data were entered in a 2 (feedback or no feedback) $\times$ 2 (position of the first test) $\times$ 2 (spacing condition: expanding or equal) ANOVA. This analysis showed only a main effect of feedback [$F(1,60) = 59.65$, $\eta_p^2 = .50$]. No other effect or interaction reached significance (all Fs < 2).

Recall order. Recall order was examined again to explore the strategies that subjects used to recall the texts. The recall order data from Experiment 2 are plotted in Figure 1 along with the data from Experiment 1. Again the data were collapsed across the four texts, the repeated tests, and the spacing conditions. The figure shows that the subjects recalled ideas in forward serial order. The overall average Pearson correlation between recall output order and input order was $r = .81$. The recall order results wholly replicated the results of Experiment 1. Subjects did not recall the most recent ideas first but instead recalled the ideas in serial order to the best of their abilities.

Discussion

The critical finding from Experiment 2 was that there was no difference between expanding and equally spaced schedules of retrieval practice. There was a large effect of repeated testing on learning, especially when the learn-

Table 4
Mean Proportions of Idea Units Recalled
on the Final Test in Experiment 2

| | Feedback Condition | | | |
| | No Feedback | | Feedback | |
Learning Condition	M	SE	M	SE
Immediate First Test				
Study once	.27	.04	.15	.05
Single test (0)	.37	.04	.34	.04
Expanding (0-1-2-3)	.55	.06	.80	.03
Equal (0-2-2-2)	.53	.06	.78	.04
Delayed First Test				
Study once	.17	.03	.24	.05
Single test (2)	.32	.05	.46	.03
Expanding (2-1-2-3)	.45	.05	.84	.03
Equal (2-2-2-2)	.52	.04	.83	.03

ing period involved feedback. Indeed, repeated retrieval of the texts with feedback produced a greater than 400% improvement in long-term retention relative to just studying the texts one time. But the particular way the repeated tests were spaced did not impact retention.

GENERAL DISCUSSION

The purpose of these experiments was to see whether expanding retrieval was a superior spaced retrieval schedule for learning expository texts relative to a schedule of equally spaced retrievals. Expanding and equally spaced retrieval schedules have been examined with paired-associate materials but not with texts—and moreover, the results with paired-associate tasks have been equivocal (see Balota et al., 2007). The present experiments showed four important findings. First, there was a testing effect: Taking a single free recall test produced better long-term retention than did reading the text and not recalling it—even when there was no feedback after the test. Second, repeated retrieval practice enhanced retention relative to practicing retrieval once on a single test. Third, rereading the texts after recalling them (and thus receiving feedback after the tests) improved learning and long-term retention. Finally, and most important for the present purposes, whether repeated retrieval attempts were expanding or equally spaced did not matter for long-term retention.

An important issue in any examination of spaced retrieval practice is that the typical comparison of expanding and equally spaced schedules confounds the position of the first test and the spacing of repeated tests. Expanding schedules typically involve an immediate first test, and equally spaced schedules typically involve a delayed first test. Any difference between expanding and equally spaced practice could be due to one or the other factor or perhaps to some interaction between the factors. In the two experiments reported here we sought to separate these two factors. In Experiment 1 we found that taking an immediate first test improved long-term retention, probably because of their increased retrieval success on that test. However, we did not find this pattern in the no-feedback conditions in Experiment 2. In fact, in the feedback conditions we found that a delayed first test produced better retention than an immediate first test. But most importantly, in both experiments there were no meaningful differences between expanding and equally spaced schedules of repeated testing.

To explain our previous findings with paired-associate materials (Karpicke & Roediger, 2007a) we proposed that expanding retrieval might not be effective because of the immediate (massed) testing used in this condition. Testing an individual item immediately after it was studied in a paired-associate learning task would prompt subjects to recall the item when it still resided in their immediate awareness, akin to rehearsing an item in massed fashion, a technique known to have minimal effects on later recall (e.g., Craik & Watkins, 1973). In the present experiments the analysis of subjects' retrieval strategies showed that when subjects recalled texts, they consistently started recalling ideas from the beginning of the texts and recalled in serial order to the best

of their abilities. This is not surprising, and we expected the result on the basis of observations in prior experiments (Roediger & Karpicke, 2006b). But the pattern with text materials differs from the typical finding—from 50 years of research on free recall of lists—that subjects recall the last items in a list first (see Deese & Kaufman, 1957; Tulving, 1968). Even on a test positioned immediately after study, none of subjects' recall of texts represented immediate retrieval of the most recent ideas from the texts. Even a single, nominally massed test can enhance later retention, as long as recall on the test does not represent retrieval of items residing in immediate memory.

Our conclusion therefore is that the particular spacing schedule of repeated tests does not produce a large impact on long-term retention. This conclusion is not consistent with Landauer and Bjork's (1978) original idea but it is supported by the bulk of recent research. For example, Carpenter and DeLosh (2005) did not find any difference between expanding and equally spaced schedules on a criterial test immediately after learning. Pyc and Rawson (2007) did not find any difference on a criterial test 40 min after learning. Logan and Balota (2008) examined the effects of several different spacing schedules with both younger and older adults and found no effect—or sometimes even an advantage of equal interval spacing—on retention 1 day after learning (see too Balota, Duchek, Sergent-Marshall, & Roediger, 2006). As long as repeated retrieval attempts are spaced rather than massed, the particular schedule of the repeated tests seems not to matter—at least under the conditions examined in the literature so far. And as noted above, it may be the case that equally spaced practice schedules will enhance long-term retention because of the delayed first retrieval attempt (Karpicke & Roediger, 2007a; Logan & Balota, 2008).

One could argue that differences between expanding and equally spaced schedules might emerge if paradigms with different features were used. For example, in our studies performance on immediate tests (without feedback) does not change much across the set of initial tests given after study. Perhaps, if conditions were arranged such that rapid intertest forgetting occurred between the repeated tests, an expanding retrieval schedule may produce better performance than equally spaced retrieval (for a similar idea see Storm, Bjork, & Storm, in press). This situation may arise if the retrieval attempts were spaced over longer delays, perhaps days apart (cf. Rawson & Kintsch, 2005), which would encourage intertest forgetting. Of course, these are empirical questions that have not yet been sufficiently examined. Nonetheless, the spacing intervals that we used in the present experiments were rather long relative to those used in typical spaced retrieval experiments with paired-associate materials. For example, in Landauer and Bjork's (1978) Experiment 2, a lag of eight trials with a presentation rate of 15 sec/trial resulted in a total lag of 2 min. In the present experiments a lag of just two periods at a rate of 3 or 4 min/period resulted in total lags of 6 or 8 min (in Experiments 1 and 2, respectively).

The theory about why expanding retrieval should be superior to equally spaced retrieval practice is twofold. First, an early initial retrieval attempt increases the likelihood

of success before forgetting has a chance to occur. Second, gradually increasing the interval between repeated retrieval attempts increases retrieval difficulty on those attempts and thereby improves later retention (Bjork, 1999). However, at least under our conditions, it is not clear that retrieval actually grows increasingly difficult across repeated tests regardless of whether the tests are expanding or equally spaced. Assuming that recall latency is a viable index of retrieval difficulty (Benjamin, Bjork, & Schwartz, 1998; Koriat & Ma'ayan, 2005), in prior research we found that recall times grew faster across repeated tests, indicating that repeated retrieval grew easier, not more difficult (Karpicke & Roediger, 2007a). If retrieval grew more difficult, we might also expect recall performances to decline across tests, but the present data show that recall remained relatively constant or even exhibited a modest degree of hypermnesia (see too Logan & Balota, 2008). Therefore it may be true that increasing the difficulty of retrieval attempts does in fact promote long-term retention (Pyc & Rawson, 2009), but the data imply that current procedures for implementing expanding retrieval do not successfully induce increasingly difficult retrieval across repeated tests. Although future research may show differences between expanding and equally spaced retrieval schedules with much longer intervals between tests, current research using relatively modest intervals has failed to obtain any consistent advantage of expanding retrieval.

CONCLUSION

The conclusions based on our experiments are clear. Retrieval practice is a powerful way to enhance long-term retention of educationally relevant texts, and repeated retrieval attempts confer benefits beyond recalling a single time. The positive effects of retrieval practice are especially great when students receive feedback after retrieval attempts (by rereading texts after attempting recall). However, there do not appear to be meaningful differences between expanding and equally spaced schedules of retrieval practice under the conditions we used. As long as retrieval attempts are spaced and not massed together, the particular way the retrieval attempts are spaced during a learning period does not matter for long-term retention.

AUTHOR NOTE

REFERENCES

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

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. (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.

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

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III (2008). Correcting a metacognitive error: Feedback increases retention of low confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **34**, 918-928.

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

Chan, J. C. K., McDermott, K. B., & Roediger, H. L., III (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 & Verbal Behavior*, **9**, 143-148.

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

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.

Deese, J., & Kaufman, R. A. (1957). Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of Experimental Psychology*, **54**, 180-187.

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 & Verbal Behavior*, **15**, 1-16.

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 371-377.

Greene, R. L. (2008). Repetition and spacing effects. In H. L. Roediger III (Ed.), *Learning and memory: A comprehensive reference. Vol. 2: Cognitive psychology of memory* (pp. 65-78). Oxford: Elsevier.

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

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

Karpicke, J. D., & Roediger, H. L., III (2007a). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **33**, 704-719.

Karpicke, J. D., & Roediger, H. L., III (2007b). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory & Language*, **57**, 151-162.

Karpicke, J. D., & Roediger, H. L., III (2008). The critical importance of retrieval for learning. *Science*, **319**, 966-968.

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

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In 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. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, & Cognition*, **15**, 257-280.

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

PASHLER, H., CEPEDA, N. J., WIXTED, J. T., & ROHRER, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 3-8.

PYC, M. A., & RAWSON, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition*, **35**, 1917-1927.

PYC, M. A., & RAWSON, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory & Language*, **60**, 437-447.

RAWSON, K. A., & KINTSCH, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology*, **97**, 70-80.

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

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

ROEDIGER, H. L., III, & THORPE, L. A. (1978). The role of recall time in producing hypermnesia. *Memory & Cognition*, **6**, 296-305.

ROGERS, B. (2001). *TOEFL CBT success*. Princeton, NJ: Peterson's.

ROHRER, D., & PASHLER, H. (2007). Increasing retention without increasing study time. *Current Directions in Psychological Science*, **16**, 183-186.

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

STORM, B. C., BJORK. R. A., & STORM, J. C. (in press). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. *Memory & Cognition*.

TULVING, E. (1968). Theoretical issues in free recall. In T. Dixon & D. Horton (Eds.), *Verbal behavior and general behavior theory* (pp. 2-36). Englewood Cliffs, NJ: Prentice Hall.

APPENDIX A
Overview of the Design and Counterbalancing of the Learning Phase in Experiment 1

Learning Condition	Period													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Immediate First Test														
Expanding (0-2-4)	S_A	T_A	S_B	T_B	T_A	V	T_B	V	V	T_A	V	T_B	V	
Equal (0-2-2)	S_A	T_A	S_B	T_B	T_A	V	T_B	T_A	V	T_B	V			
Delayed First Test														
Expanding (2-2-4)	S_A	S_B	V	T_A	T_B	V	T_A	T_B	V	V	V	T_A	T_B	V
Equal (2-2-2)	S_A	S_B	V	T_A	T_B	V	T_A	T_B	V	T_A	T_B	V		

Note—S, study period; T, test period; V, video game filler period. A and B denote the two passages, which were counterbalanced to create two orders per condition. The expanding immediate (0-2-4) and equal delayed (2-2-2) conditions represent the conditions typically used in expanding retrieval experiments.

APPENDIX B
Overview of the Design and Counterbalancing of the Learning Phase in Experiment 2

Order	Period Number																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Immediate First Test Conditions																		
1	S_A	S_B	T_B	F	F	S_C	T_C	S_D	T_D	T_C	T_D	F	T_C	T_D	F	T_C	F	T_D
2	S_A	T_A	S_B	T_B	F	F	T_B	S_C	T_C	T_B	T_C	S_D	T_B	T_C	F	T_C	F	T_C
3	S_A	T_A	F	F	T_A	S_B	T_B	T_A	T_B	S_C	T_A	T_B	F	F	F	T_B	S_D	S_D
4	S_A	T_A	S_B	T_A	S_C	T_C	T_A	S_D	T_D	F	T_A	T_D	F	F	T_D	F	F	T_D
Delayed First Test Conditions																		
1	S_A	F	F	S_B	S_C	S_D	T_B	T_C	T_D	T_B	T_D	F	T_B	T_D	F	T_B	F	T_D
2	S_A	S_B	F	T_A	T_B	S_C	T_A	S_D	T_C	T_A	T_C	F	T_A	T_C	F	F	F	T_C
3	S_A	S_B	S_C	T_A	T_B	S_D	T_B	F	T_D	T_B	F	T_D	F	T_B	T_D	F	F	T_D
4	F	S_A	F	S_B	T_A	S_C	T_A	S_D	T_C	T_A	T_D	T_C	F	T_A	T_C	F	F	T_C

Note—S, study period; T, test period; F, filler period. A, B, C, and D represent the four passages, which were always presented in the same order (*Bamboo*, *Diving*, *Eugene O'Neill*, and *Vanuatu*, respectively).