Psychon Bull Rev
DOI 10.3758/s13423-017-1366-9

BRIEF REPORT

Diminishing-cues retrieval practice: A memory-enhancing technique that works when regular testing doesn't

Joshua L. Fiechter[1] · Aaron S. Benjamin[1]

Abstract Retrieval practice has been shown to be a highly effective tool for enhancing memory, a fact that has led to major changes to educational practice and technology. However, when initial learning is poor, initial retrieval practice is unlikely to be successful and long-term benefits of retrieval practice are compromised or nonexistent. Here, we investigate the benefit of a scaffolded retrieval technique called diminishing-cues retrieval practice (Finley, Benjamin, Hays, Bjork, & Kornell, *Journal of Memory and Language*, *64*, 289–298, 2011). Under learning conditions that favored a strong testing effect, diminishing cues and standard retrieval practice both enhanced memory performance relative to restudy. Critically, under learning conditions where standard retrieval practice was not helpful, diminishing cues enhanced memory performance substantially. These experiments demonstrate that diminishing-cues retrieval practice can widen the range of conditions under which testing can benefit memory, and so can serve as a model for the broader application of testing-based techniques for enhancing learning.

Keywords Memory · Human memory and learning · Retrieval cues and memory

When learners retrieve information from memory, their long-term retention of that information is often enhanced compared to learners that restudy the material. Evidence for the *testing effect* reaches back nearly 100 years (Gates, 1917); more recent work has renewed interest in the cognitive benefits conferred by retrieval (Roediger & Karpicke, 2006a). For the past decade, memory researchers have devoted considerable attention to the benefits of testing, finding it to be an effective encoding strategy for a wide range of materials, including single words (e.g., Carpenter & DeLosh, 2006), word pairs (e.g., Carpenter, 2009), text passages (e.g., Roediger & Karpicke, 2006b), and nonverbal materials (e.g., Carpenter & Pashler, 2007). In addition to basic laboratory tasks, testing has also been shown to be effective in educational settings, outperforming even the most highly recommended educational practices (Karpicke & Blunt, 2011). In-class quizzes enhance test performance for students ranging from middle school to college (McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; McDaniel, Anderson, Derbish, & Morissette, 2007; McDermott, Agarwal, D'Antonio, Roediger, & McDaniel, 2014; Roediger, Agarwal, McDaniel, & McDermott, 2011; Weinstein, Nunes, & Karpicke, 2016). Classification testing also enhances the ability to generalize knowledge to new category members (Jacoby, Wahlheim, & Coane, 2010). In-class quizzing similarly enhances performance on exam questions requiring an application of knowledge (McDaniel, Thomas, Agarwal, & McDermott, 2013). So promising is the potential of retrieval practice as a learning tool that memory researchers have begun advocating for a larger emphasis on testing as a means of enhancing, and not just assessing, knowledge in educational policy (Benjamin & Pashler, 2015).

Retrieval confers significant benefits on retention, but there is a noteworthy trade-off inherent to testing: Testing strongly enhances retention when retrieval practice is successful, but memory for information that is not successfully retrieved on practice tests is not enhanced. As a result, subsequent memory

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13423-017-1366-9) contains supplementary material, which is available to authorized users.

✉ Joshua L. Fiechter
fiechte2@illinois.edu

[1] Department of Psychology, University of Illinois at Urbana–Champaign, 603 E. Daniel St., Champaign, IL 61820, USA

Published online: 28 August 2017

hinges largely on retrieval practice success: on a final assessment, items successfully retrieved during practice will mostly be remembered and unsuccessfully retrieved items will mostly be forgotten (Kornell, Bjork, & Garcia, 2011). If learners are doing well during retrieval practice, this trade-off is desirable; if they are struggling, however, the costs may outweigh the benefits, revealing a major boundary condition on the benefits of standard retrieval practice. A recent meta-analysis offered evidence for such a boundary condition: In studies where performance during retrieval practice was below 50%, and in which learners did not receive item-by-item feedback on their practice performance, the testing effect is absent (Rowland, 2014).

Given that the benefits of testing are severely reduced in circumstances where initial learning is poor, and that testing is increasingly working its way into applied environments such as education, it is important to find a technique that generalizes those benefits to a wider range of learning conditions. Such a technique could incorporate scaffolding, an educational and training technique where students are provided progressively less assistance until that assistance is no longer necessary (Wood, Bruner, & Ross, 1976). Past research has found evidence for the benefits of scaffolded practice in a variety of domains. In the motor learning literature, providing progressively less feedback over the duration of practice results in better learning than consistently provided feedback (Winstein & Schmidt, 1990; Wulf & Schmidt, 1989). Work with example problems in mathematics and science instruction has found that “fading” examples, by removing steps from worked out problems, promotes learning more than having learners restudy intact examples (Atkinson, Renkl, & Merril, 2003). Similarly, mathematics examples that progress from concrete to abstract benefit learning more than concrete or abstract examples alone (Fyfe, McNeil, & Borjas, 2015; McNeil & Fyfe, 2012). In verbal learning, expanding retrieval practice, where learners retrieve information soon after studying it and then again at progressively longer lags, has been shown to enhance learning more than retrieving information at uniform lags (Cull, Shaughnessy, & Zechmeister, 1996; Landauer & Bjork, 1978; Storm, Bjork, & Storm, 2010).

Although there is evidence that scaffolding is effective for learning, it is not always beneficial to facilitate practice during the acquisition of skills or knowledge (e.g. Schmidt & Bjork, 1992). For example, numerous studies in the motor learning literature have found that practicing related tasks in a randomly determined order suppresses practice performance but leads to superior long-term performance over a blocked schedule (Lee & Magill, 1983; Shea & Morgan, 1979; see Brady, 1998, for a review). That is, blocked practice, which facilitates acquisition performance—just as a scaffolded schedule would—is in fact detrimental to long-term retention. In verbal learning, there is evidence that practicing retrieval from impoverished cues leads to better memory than practicing retrieval from more complete cues (Carpenter & DeLosh, 2006) and that retrieval practice conditions that require more effort are generally better for enhancing long-term memory than conditions that require less effort (Pyc & Rawson, 2009). Accordingly, Karpicke and Roediger (2007) found that, at longer retention intervals, expanding retrieval practice actually led to impaired long-term memory relative to uniform retrieval intervals ostensibly because the early tests in expanding retrieval practice required too little effort (see also Cull, 2000; Logan & Balota, 2008). If scaffolded retrieval practice promotes retrieval that is too shallow or too easy, then scaffolding could become a liability.

We evaluate whether retrieval practice, one of the most potent means of encoding, can be made even more beneficial by incorporating the principles of scaffolding during practice. We evaluate the merits of a technique called diminishing-cues retrieval practice: a study method where learners are exposed to progressively impoverished cues until they must retrieve target information without any additional assistance. This technique was introduced by Finley et al. (2011). In their experiments, learners practiced English–Iñupiaq word pairs (e.g., tea–*saiyu*) on three different schedules: (a) diminishing-cues retrieval practice, where learners initially saw the complete Iñupiaq word before letters were randomly omitted from the Iñupiaq words, one at a time, over six practice rounds; (b) accumulating-cues retrieval practice, where learners initially saw no letters of the Iñupiaq word before they were randomly included, one at a time, over six rounds of practice; and (c) a control restudy condition, where learners saw the entirety of each word pair during each round of practice. They found that diminishing-cues retrieval practice was superior to both restudy and accumulating-cues retrieval practice when item-by-item feedback was not presented, and that diminishing- and accumulating-cues retrieval practice were both superior to restudy, to approximately the same degree, in the presence of feedback.

Critically, Finley et al. (2011) did not evaluate the benefits of diminishing-cues retrieval practice relative to standard retrieval practice. Although the scaffolded nature of diminishing-cues retrieval practice benefited learners more than restudy in their experiments, it remains an open question as to whether it is more or less beneficial than standard retrieval practice, which has been shown to be greatly beneficial in a huge variety of circumstances (Rowland, 2014). It may be that scaffolded retrieval is specifically effective under very difficult learning conditions, where the benefits of retrieval practice are minimal or nonexistent, but not in easier conditions where standard retrieval is likely to be successful or if item-by-item feedback offsets the costs of unsuccessful retrievals. With that in mind, we created three different learning scenarios, each corresponding to an increasingly greater probability of yielding a testing effect. A testing effect was not expected in our first scenario, in which initial retrievability was less than 50% and feedback was not provided (Hedges' g = 0.03, as reported in Rowland, 2014). A testing effect was more likely in our second scenario, in which initial retrievability was greater than 50% and feedback was still not provided (g = .29). Finally, a testing effect was considered quite probable in our third scenario, where initial retrievability was less than 50% but learners were provided item-by-item feedback (g = .99).

To create our three learning conditions, we varied our experiments along several dimensions. We used three stimuli sets (English–Iñupiaq word pairs, low-association English word pairs, and Swahili–English word pairs), varied the number of study rounds (1–3) and practice rounds (1–6), varied the retention interval (10 minutes, 24 hours, and 1 week), and also varied our implementation of the diminishing cues schedule. The particular details of each experiment are outlined in the method sections; a summary of the experiments can be found in Table 1. The performance data from individual experiments can be found in Table S1 in the online supplementary materials; results summarized across similar conditions are presented here, in the main text.

Experiments 1a–1e

Method

Subjects Subjects in all experiments were recruited via Amazon's Mechanical Turk service. All experiments recruited approximately 60 subjects. (For two-part experiments, we initially recruited 80 participants and, based on return rates for previous online studies of ours, expected approximately 75% of those participants to return for the second part; if fewer than 60 participants ultimately returned for the second part, we then recruited additional subjects as needed until we gathered data from at least 60 people.) This particular sample size was obtained from a power analysis that was conducted with the aim of replicating, with 80% power, Finley et al.'s (2011) observed effect size between diminishing-cues retrieval practice and restudy. Sixty subjects completed Experiment 1a, 60 subjects completed Experiment 1b, 66 subjects completed Experiment 1c, 68 subjects completed Experiment 1d, and 64 subjects completed Experiment 1e. The median age for all participants across our first five experiments, including those that only completed the first part of a two-part experiment, was 36.5 years; the age range was 21 to 76 years.

Table 1 Summary of the experiments in terms of number of study rounds, number of practice rounds, provision of feedback, retention interval, and sample size

Experiment	# Study	# Practice	Feedback	RI	N
1a[1]	3	6	No	10 min	60
1b[1]	3	6	No	24 hours	60
1c[1]	3	6	No	1 week	66
1d[1]	3	3	No	24 hours	68
1e[2]	1	1	No	24 hours	64
2a[3]	2	2	No	24 hours	66
2b[3]	1 or 3	2	No	24 hours	184
3a[1]	1	3	Yes	24 hours	66
3b[1]	1	3	Yes	24 hours	60
3c[2]	1	1	Yes	24 hours	63

[1] English–Iñupiaq word pairs

[2] Low-association English word pairs

[3] Swahili–English word pairs

Design Each experiment used a three-level (practice method: restudy vs. retrieval practice vs. diminishing-cues retrieval practice) within-subjects design.

Materials Stimuli in Experiments 1a–1d were 12 English–Iñupiaq word pairs (e.g., tea–*saiyu*). All Iñupiaq targets were five letters long. Stimuli in Experiment 1e were 60 low-association English pairs (e.g., chart–statistics), 36 of which were randomly selected for each subject. Tables S2 and S3 in the Supplemental Material present the stimuli used in these experiments.

Procedure All experiments shared the same general methodology: a study phase, followed by a 60-second go/no-go distractor task, a practice phase, a retention interval, and a final test phase. During the study phase, subjects were presented with a word pair for 4 seconds. Most of our experiments had multiple study rounds (explained in more detail below). Following the distractor task, subjects practiced on all three schedules of interest, which were interleaved randomly: in the restudy condition, subjects were presented with a complete word pair and asked to type in the target word; in the retrieval practice condition, subjects were presented with a cue word and asked to either provide the target or type in a question mark (?) if they were uncertain; in the diminishing cues condition, subjects were shown a cue word and a portion of the target word and asked to provide the target or type in a "?" if they were uncertain. Letters from the target word were dropped over subsequent rounds of practice—for example, in our first three experiments, subjects saw a complete target word on the first practice round (*saiyu*), then in the next round the same target with one additional letter randomly omitted (*s_iyu*), and so on, until they had to retrieve the complete target word. (See Fig. 1a for an illustration of the practice phase in Experiments 1a–1c.) The test phase consisted of a randomly ordered presentation of cue words, and subjects were asked to provide the targets.

We conducted Experiments 1a–1c with the goal of constructing a forgetting curve for our three practice conditions

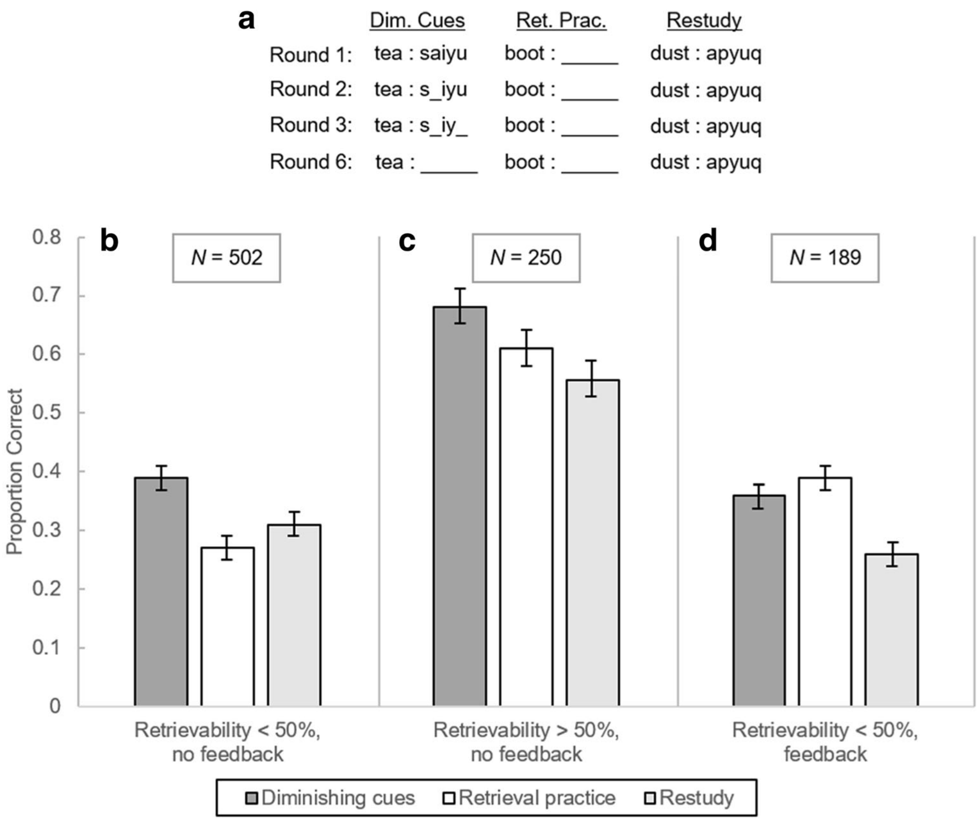

Fig. 1 Practice phase schematic and results of experiments. Schematic of practice phase for Experiments 1a–1c (*a*). Proportion of words correctly recalled in Experiments 1a–1e (and one condition from Experiment 2b) (**b**), Experiments 2a–2b (**c**), and Experiments 3a–3c (**d**). Height of *error bars* indicates within-subjects 95% confidence interval across conditions (Loftus & Masson, 1994)

over retention intervals of 10 minutes, 24 hours, and 1 week. In the study phase for these experiments, the pairs were randomly arranged; subjects studied the items three times, with each round of study maintaining the same random order. During the practice phase, four items were randomly assigned to each of the practice conditions. Practice followed a predetermined schedule that was designed under the initial constraint that no more than two items from each condition appear in each half of the initial practice round. Subsequent practice rounds had the same constraint; furthermore, the presentation order of items practiced in each half of the initial practice round was preserved in subsequent practice rounds. For items in the diminishing cues condition, subjects initially saw the complete word pair, with one letter randomly dropped from the target word over subsequent rounds until no more letters were shown (requiring six rounds of practice). After the practice phase, subjects in our first three experiments waited either 10 minutes (Experiment 1a), 24 hours (1b) or 1 week (1c) before taking a final test. (All subsequent experiments used a 24-hour retention interval.)

We conducted Experiment 1d to evaluate the benefits of diminishing-cues retrieval practice with fewer rounds of practice. Experiment 1d used the same general procedure as Experiments 1a–1c. However, we provided participants with three practice rounds instead of six, and so we changed the diminishing cues condition to drop one, three, and five letters at each of the three practice rounds (as opposed to the letter-by-letter dropping procedure used in Experiments 1a–1c, which required six rounds to implement).

We conducted Experiment 1e to evaluate the benefits of diminishing-cues retrieval practice with just a single round of practice. The English–Iñupiaq word pairs that we had used to this point were too difficult to remember after a single practice round; we therefore switched our stimuli to 60 low-association English word pairs (taken from Hays, 2009). Of the 60 pairs, 36 were selected at random for each subject. All target words were 8 to 10 letters long. Subjects saw the items only once during the study phase. The practice phase consisted of a single round of practice; items in the diminishing cues condition had five letters randomly dropped from the target word. Practice presentation was randomly determined, with the constraint that four words from each condition appeared in each third of practice.

Results

This series of experiments compared the benefits of diminishing cues and retrieval practice under conditions where standard retrieval practice was expected to confer relatively little benefit to memory. Testing is rarely beneficial when two criteria are met: (1) initial retrievability—that is, retrieval practice performance in the first practice round—of items during the practice phase is less than 50%, and (2) learners are deprived of feedback during the practice phase. To control for experiment effects in our combined data sets, we fit mixed-effects models to our data that included a fixed effect of condition and random intercepts for experiments and participants. Subsequent analyses of group differences used the variance estimates from these models. Because we anticipated nondifferences between some groups, and because these null results would be meaningful and interpretable, we analyzed data using Bayesian *t* tests (Rouder, Speckman, Sun, Morey, & Iverson, 2009). Bayesian analyses have the advantage of evaluating evidence for both the null and alternative hypotheses (Gallistel, 2009). Comparisons involving the restudy condition use a one-tailed test because performance in the restudy condition was never expected to be better than in either of the other two conditions.

The compiled data from Experiments 1a–1e are plotted in Fig. 1b. (We replicated these findings in one condition from Experiment 2b; those data are also included in the plot.) As expected, we found no advantage for retrieval practice (M = 0.27, SD = 0.34) over restudy (M = 0.31, SD = 0.36), BF_{10} = 0.02. Critically, we found highly convincing evidence that diminishing-cues retrieval practice (M = 0.39, SD = 0.37) was superior to standard retrieval practice ($\mathrm{BF}_{10} = 1.06 \times 10^{12}$, d = 0.51), and also to restudy ($\mathrm{BF}_{10} = 1.07 \times 10^{6}$, d = 0.37).

Discussion

These results provide very strong evidence that scaffolded retrieval allows learners to reap the benefits of retrieval practice without incurring the costs of failed retrieval attempts: across these conditions, subjects remembered 44% more information when they studied using diminishing-cues retrieval practice instead of standard retrieval practice. We next evaluated diminishing cues under conditions in which retrieval practice is more effective. If partial cues compromise the value of retrieval by rendering some of the retrieval events too easy, then under these conditions there should be an advantage for standard retrieval practice. We sought a stronger testing effect by increasing initial retrievability to higher than 50%. Our next set of experiments evaluated the benefits of diminishing-cues retrieval practice when we could expect a larger testing effect.

Experiments 2a–2b

Method

Subjects Sixty-six subjects completed Experiment 2a and 184 subjects completed Experiment 2b. The larger sample size in Experiment 2b had two motivations. First, the larger sample was based partly on a power analysis for replicating, with 80% power, a small testing effect observed in Experiment 2a. Second, we wanted to increase the number of item-level observations per condition, since Experiment 2b used the same number of to-be-learned items as Experiment 2a but had twice as many within-subject conditions. The median age for all participants across these next two experiments, including those that only completed the first of two parts, was 36 years; the age range was 18 to 83 years.

Design Experiment 2a used a three-level (practice method: restudy vs. retrieval practice vs. diminishing-cues retrieval practice) within-subjects design. Experiment 2b had an additional manipulation of initial study exposure, resulting in a 2 (number of study exposures: 1 vs. 3) × 3 (practice method: restudy vs. retrieval practice vs. diminishing-cues retrieval practice) within-subjects design.

Materials Stimuli in these experiments were 12 Swahili–English word pairs (e.g., *malkia*–queen). All English targets were five letters long. Table S4 in the Supplemental Material presents the stimuli used in these experiments.

Procedure We conducted Experiment 2a to evaluate our three practice conditions with easier stimuli, that is, stimuli whose initial retrievability were higher than 50%. Both the English–Iñupiaq word pairs and the low-association English pairs that we had previously used were too difficult to produce our desired higher retrievability; we therefore switched to Swahili–English word pairs. English targets should be easier to learn than Iñupiaq targets, and we restricted ourselves to pairs that had five-letter English targets, which should be easier to rehearse than the longer, 8 to 10 letter targets in the low-association English pairs (Baddeley, Thomson, & Buchanan, 1975). Using the normed pairs from Nelson & Dunlosky (1994), we selected 12 pairs with five-letter targets that had the highest first-trial recall (using second-trial recall in the case of ties). To avoid ceiling performance, we determined, through pilot testing, to use two study presentations and two practice rounds. Each study round presented the items in a newly randomized order; each practice round used the same presentation constraints as in Experiments 1a–1d. In the diminishing cues condition, target words were seen with two letters randomly dropped during the first round, and all letters dropped during the second round.

Experiment 2a revealed a different pattern of results than our previous experiments, with diminishing-cues retrieval practice benefitting memory about the same as retrieval practice relative to restudy (see Table S1 in the online supplementary materials). However, our findings for this experiment were not convincing and so we conducted Experiment 2b to replicate the pattern of findings from Experiment 2a using a higher-powered design. A second goal of Experiment 2b was to elicit the two patterns of results that we had observed up to this point (from Experiments 1a–1e and Experiment 2a) in a single experiment. For this second goal, we included a manipulation of initial retrievability, which we believed to be the source of our two patterns of findings. The stimuli in Experiment 2b were again the 12 Swahili–English word pairs, but now half of the items were assigned to be studied three times and the other half to be studied only once. Study was arranged into three rounds, with each round consisting of all the thrice-studied items and two randomly selected once-studied items. All rounds followed the same randomly determined order, with different once-studied items occupying the same position over rounds. Practice for the diminishing cues condition was the same as in Experiment 2a.

Results

The combined data from Experiments 2a and 2b are presented in Fig. 1c. (The condition from Experiment 2b that is included in Fig. 1b is not included in Fig. 1c.) Subjects indeed found these tasks to be easier: initial retrievability was 66%, as opposed to 25% in Experiments 1a–1e. Unlike our initial experiments, retrieval practice (M = 0.61, SD = 0.39) was now numerically better than restudy (M = 0.56, SD = 0.38), though evidence for a testing effect was weak, BF_{10} = 1.94, d = 0.20. Importantly, diminishing-cues retrieval practice (M = 0.68, SD = 0.36) was superior to both standard retrieval practice (BF_{10} = 6.18, d = 0.28) and to restudy (BF_{10} = 94,426, d = 0.48).

Discussion

After enhancing the retrievability of items (and consequently observing higher performance in the standard retrieval practice condition) we again found that diminishing-cues retrieval practice was the superior study method. However, the benefits of testing were still small, leaving open the possibility that the benefits of standard retrieval practice might be greater than those of diminishing-cues retrieval practice in conditions that are most favorable to yielding large testing effects. In the next set of experiments, we utilized conditions where standard retrieval practice is at its most potent: when initial retrieval success is low (i.e., less than 50%) but practice is accompanied by item-by-item feedback of the correct response during the practice phase. Our final set of experiments used the same items as our first—to ensure low retrievability—but we now provided participants with feedback to foster a strong testing effect.

Experiments 3a–3c

Method

Subjects Sixty-six subjects completed Experiment 3a, 60 subjects completed Experiment 3b, and 63 subjects completed Experiment 3c. The median age for all participants across our final three experiments, including those that only completed the first of two parts, was 38 years; the age range was 20 to 71 years.

Design All experiments used a three-level (practice method: restudy vs. retrieval practice vs. diminishing-cues retrieval practice) within-subjects design.

Materials Stimuli in Experiments 3a and 3b were English–Iñupiaq word pairs. Stimuli in Experiment 3c were 60 low-association English word pairs, 36 of which were randomly selected for each participant.

Procedure We conducted Experiment 3a to see how our three practice conditions fared under provision of feedback. Because feedback made the task substantially easier, we went back to the difficult English–Iñupiaq word pairs that we used in Experiments 1a–1d. Even with these difficult stimuli, we were concerned that providing three study exposures and six practice rounds (as we did in Experiments 1a–1c) would lead to ceiling effects on retention, and, after pilot testing, decided on one study exposure and three rounds of practice (with the diminishing cues schedule dropping one, three, and five letters over the course of practice, just as in Experiment 1d). Feedback in our experiments worked as follows: In the practice phase, following an attempted retrieval of the target word, subjects saw a 4-second presentation of a complete word pair corresponding to the pair they had just practiced. In the restudy condition, feedback was tantamount to 4 additional seconds of study. Subjects received feedback in the diminishing cues and retrieval practice conditions regardless of whether their response was correct or incorrect.

We conducted Experiment 3b to evaluate whether the benefits of the diminishing cues schedule obtained if participants were never required to retrieve the entire target word. We adjusted the diminishing cues condition such that target words were presented with one, two, and three letters dropped at random (as opposed to one, three, and five letters in Experiment 3a).

As was the case with Experiment 1e, we conducted Experiment 3c to see how diminishing cues fared with a single round of practice. We used the set of 60 low-association English word pairs that we also used in Experiment 1e.

Results

The compiled data from Experiments 3a–3c are plotted in Fig. 1d. As expected, items in the retrieval practice condition (M = 0.39, SD = 0.32) were substantially better remembered than items in the restudy condition (M= 0.26, SD = 0.31), $BF_{10} = 1.05 \times 10^{10}$, d = 0.78. Unlike our previous experiments, learners now showed a strong testing effect. Diminishing-cues retrieval practice (M = 0.36, SD = 0.32) was also superior to the restudy condition, $BF_{10} = 9.85 \times 10^{5}$, d = 0.64, and—critically—the data suggest that diminishing-cues retrieval practice was just as effective as standard retrieval practice, $BF_{10} = 0.35$.

Discussion

Our final series of experiments demonstrated a strong testing effect: Subjects remembered 50% more information if they practiced retrieving items rather than restudying. And, of primary interest, diminishing-cues retrieval practice was just as effective for memory as retrieval practice. Thus, under conditions that produce the strongest testing effects, we found that diminishing-cues retrieval practice was equally beneficial. Diminishing-cues retrieval practice appears to generalize to more learning situations than does standard retrieval practice.

General discussion

Across three sets of experiments, our findings are twofold. First, diminishing-cues retrieval practice works when testing does not. Across the experiments in which we found little or no testing effect, we did find an advantage of diminishing-cues retrieval practice over restudy and over retrieval practice. It appears that, when a task is sufficiently difficult such that retrieval of items during practice is unlikely, learners benefit from the accumulation of retrieval demands that grow over the course of practice. Such a schedule increases the probability of successful retrieval while still maintaining sufficient challenge. In contrast, standard retrieval practice is unlikely to yield memory benefits when initial retrieval is too difficult, and learning suffers as a result.

Second, when testing effects are at their strongest, diminishing-cues retrieval practice is equally effective. In experiments in which feedback was provided, standard retrieval practice and diminishing-cues retrieval practice were both superior to restudy. Diminishing-cues retrieval practice—in addition to scaffolding retrieval demands—may also provide some form of feedback to learners; consequently, provision of feedback brings standard retrieval practice up to the level of diminishing-cues retrieval practice. Our data support this interpretation: Performance in the diminishing-cues retrieval practice condition was relatively unchanged by feedback (36% vs. 39% in the feedback and no feedback experiments, respectively), while performance in the retrieval practice condition was substantially better in the feedback experiments (39% vs. 23%).

Other explanations beyond the feedback provided by diminishing cues are also possible. The two-stage model of retrieval (Kornell, Klein, & Rawson, 2015) posits that the benefits of retrieval arise from (1) attempting to retrieve information and (2) having exposure to the correct information. This assertion implies that retrieval success is most valuable in the absence of feedback, where successful retrieval is the only means of accessing the correct information. Accordingly, participants in our experiments benefited from diminishing-cues retrieval practice, which fosters retrieval success, when feedback was not provided. However, when we did provide feedback, the benefits of standard and diminishing-cues retrieval practice were the same. That is, the enhanced retrieval success fostered by the diminishing cues schedule was offset by the equally beneficial opportunity to receive feedback—particularly after a failed retrieval—in the standard retrieval practice condition.

While we sought to enhance the generalizability of our findings by creating a variety of learning conditions, our experiments do face limitations when considering how our study might extend to learning outside the laboratory. First, all our experiments compared practice conditions within subject. Although within-subject designs are desirable for their increased power, a learner outside the laboratory would almost certainly engage a single encoding strategy, and a between-subject manipulation may more closely mimic learners' study habits. Furthermore, according to Rowland's (2014) data, testing effects are larger in between-subject (g = 0.69) than within-subject (g = 0.43) designs; whether the benefits of diminishing-cues retrieval practice are also larger in between-subject designs remains to be seen. Second, the stimuli that we used were substantially less complex than what would be learned, for example, as part of a college course (with the caveat that foreign language words and their translations are frequently studied in second language courses). To once again refer to Rowland's data, testing effects are just as large with more complex prose materials (g = 0.58) as they are with simpler word pairs (g = 0.59); thus, more complex materials may enjoy the benefits of diminishing-cues retrieval practice to a similar extent as do simpler stimuli, though this claim has not been evaluated. Future research should address these concerns by implementing diminishing-cues retrieval practice in scenarios that more closely approximate learning in applied environments.

Another feature of our experiments that merits discussion is the wide age range of our participants: Across all experiments, we collected data from individuals ranging from 18 to 83 years old. We avoided linking participants' demographic information to their experiment data; we are thus uncertain of how participants' ages may have interacted with our three practice conditions. Past research on the benefits of retrieval practice versus restudy across age groups has found that testing effects are equally large for older and younger adults (Coane, 2013; Meyer & Logan, 2013). In a comparison of the benefits of expanding versus uniformly spaced retrieval practice, Logan & Balota (2008) found that older adults benefited more (on a same-day test) and suffered fewer costs (after a 24-hour delay) from expanding retrieval practice than did younger adults. Insomuch as expanding retrieval practice approximates our diminishing cues technique, these data suggest that diminishing-cues retrieval practice may have benefited our older subjects more than our younger subjects. Despite these potential age-related differences, a vast minority of our subjects qualified as older adults (only 11% of our participants reported being older than 55 years; of these people, not all may have completed both sessions of our two-part experiments) and so we are confident that the benefits of diminishing cues extend to learners of all ages.

Diminishing-cues retrieval practice appears to provide a more generally effective means of implementing retrieval practice—one that works across a wider range of the task difficulty spectrum. Scaffolding, in addition to being heralded as an instructional technique, is also effective as a means of guiding retrieval practice. In learning situations in which graduated retrieval difficulty is possible, it may provide a superior means of ensuring long-term retention than standard retrieval practice.

Data availability

Our complete data set and an accompanying R script are available online at the Open Science Framework (https://osf.io/xztfb/).

References

Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. *Journal of Educational Psychology*, *95*, 774–783.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *14*, 575–589.

Benjamin, A. S., & Pashler, H. (2015). The value of standardized testing: A perspective from cognitive psychology. *Policy Insights from the Behavioral and Brain Sciences*, *2*, 13–23.

Brady, F. (1998). A theoretical and empirical review of the contextual interference effect and the learning of motor skills. *Quest*, *50*, 266–293.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1563–1569.

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

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review*, *14*, 474–478.

Coane, J. H. (2013). Retrieval practice and elaborative encoding benefit memory in younger and older adults. *Journal of Applied Research in Memory and Cognition*, *2*, 95–100.

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.

Finley, J. R., Benjamin, A. S., Hays, M. J., Bjork, R. A., & Kornell, N. (2011). Benefits of accumulating versus diminishing cues in recall. *Journal of Memory and Language*, *64*, 289–298.

Fyfe, E. R., McNeil, N. M., & Borjas, S. (2015). Benefits of "concreteness fading" for children's mathematics understanding. *Learning and Instruction*, *35*, 104–120.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439–453.

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

Hays, M. J. (2009). Using adaptive feedback to optimize learning (Unpublished doctoral dissertation). University of California, Los Angeles, CA.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1441–1451.

Karpicke, J. D. & Blunt, J. R. (2011). Retrieval practices produces learning than elaborative studying with concept mapping. *Science*, *331*, 772–775.

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

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*, 85–97.

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 283–294.

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.

Lee, T. D. & Magill, R. A. (1983). The locus of contextual interference in motor-skill acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 730–746.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*, 476–490.

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, and Cognition*, *15*, 257–280.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*, 399–414.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morissette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*, 494–513.

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology*, *27*, 360–372.

McDermott, K. B., Agarwal, P. K., D'Antonio, Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, *20*, 3–21.

McNeil, N. M. & Fyfe, E. R. (2012). "Concreteness fading" promotes transfer of mathematical knowledge. *Learning and Instruction*, *22*, 440–448.

Meyer, A. N. D., & Logan, J. M. (2013). Taking the testing effect beyond the college freshman: Benefits for lifelong learning. *Psychology and Aging*, *28*, 142–147.

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory*, *2*, 325–335.

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

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*, 382–395.

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.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*, 1432–1463.

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.

Shea, J. B., & Morgan, R. L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory*, *5*, 179–187.

Storm, B. C., Bjork, R. A. & Storm, J. C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. *Memory & Cognition*, *38*, 244–253.

Weinstein, Y., Nunes, L. D., & Karpicke, J. D. (2016). On the placement of practice questions during study. *Journal of Experimental Psychology: Applied*, *22*, 72–84.

Winstein, C. J., & Schmidt, R. A. (1990). Reduced frequency of knowledge of results enhances motor skill learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 677–691.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*, 89–100.

Wulf, G., & Schmidt, R. A. (1989). The learning of generalized motor programs: Reducing the relative frequency of knowledge of results enhances memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 748–757.