Psychonomic Bulletin & Review
https://doi.org/10.3758/s13423-019-01617-6

# Techniques for scaffolding retrieval practice: The costs and benefits of adaptive versus diminishing cues

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



## Abstract

Testing is a powerful enhancer of memory. However, if initial encoding is poor, and subsequent retrieval practice is likely to fail, then the benefits of testing are diminished or even eliminated. Previous work has suggested that the benefits of testing may be preserved under difficult conditions with a scaffolded technique called diminishing-cues retrieval practice (DCRP; Fiechter & Benjamin, *Psychonomic Bulletin & Review, 25*(5), 1868–1876, 2018). DCRP provides increasing retrieval demands over practice, but does not adapt to individual learners or to materials of varying difficulty. Here, we evaluate a new technique called adaptive-cues retrieval practice (ACRP). ACRP adapts to an individual's moment-to-moment ability by providing within-trial accumulated cuing, generating more demanding retrieval practice for better learned items. Across six experiments, learners practiced English–Iñupiaq word pairs using ACRP, standard retrieval practice, restudy, and DCRP. ACRP is even more effective than DCRP in situations where standard retrieval practice is ineffective. When testing is most effective, ACRP, DCRP, and standard retrieval practice all enhance memory to approximately the same degree, but DCRP requires the least practice time. Our findings suggest that DCRP is a more efficient technique for learning, but that the benefits of ACRP extend to more learning scenarios than those of any other identified practice regimen.

Few phenomena in the memory literature have received as much scrutiny as the *testing effect*, or the finding that retrieving information enhances the likelihood of that information being retrieved later (e.g., Abbot, 1909). Over a variety of learning situations and with myriad stimuli, the benefits of retrieval practice have proven to be remarkably robust (see Rowland, 2014, for a review). And yet, for all its benefits, there are circumstances under which testing is not helpful, or is even costly. In particular, the benefits of testing are greatly diminished when items are difficult to learn and corrective feedback is absent (Rowland, 2014). That is, retrieval only benefits those items that are successfully retrieved (Kornell, Bjork, & Garcia, 2011; see also Benjamin & Tullis, 2010); unretrieved items receive no strengthening from testing. When a sufficiently small number of items are successfully retrieved during practice, and no corrective feedback is provided, the memory benefits that accrue to the successfully retrieved subset are insufficient to outweigh the cost to unretrieved items.

A similar scenario plays out in the *spacing effect*: Memory tends to be enhanced by stimulus presentations that are spaced apart rather than massed (e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006), but the benefits of spacing are no longer apparent when items are spaced too far apart (e.g., Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008). Benjamin and Tullis (2010) explained these joint effects of spacing in terms of *remindings*: Learners must be reminded of an initial presentation upon viewing a repetition; overly long lags diminish the prospects of reminding, and so the benefits of spacing diminish. Furthermore, memory is also less likely to be strengthened when reminding is too easy, as in the case of massed repetitions. Thus, reminding theory argues that there is a sweet spot along the retrieval-depth continuum that balances the trade-off between successful reminding and mnemonic benefits.

Of key importance to the present discussion is that the tenets of reminding theory provide a comprehensive account of the testing effect when feedback is not provided. Specifically, just as reminding theory posits that reminding should be neither too difficult nor too easy, past studies on the testing effect have demonstrated that retrieval is less

**Electronic supplementary material** The online version of this article (https://doi.org/10.3758/s13423-019-01617-6) contains supplementary material, which is available to authorized users.

✉ Joshua L. Fiechter
josh.fiechter@gmail.com

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

Published online: 03 June 2019

beneficial if it demands too little processing (Carpenter & DeLosh, 2006; Pyc & Rawson, 2009), and also that it is unhelpful if it demands too much (Jang, Wixted, Pecher, Zeelenberg, & Huber, 2012; Rowland, 2014). The relationship between retrieval difficulty and aggregate memory strengthening is therefore an inverted U—tests that are too difficult or too easy offer suboptimal memory enhancement (see Fig. 1). This account of retrieval practice is the first to capture the dual effects of retrieval difficulty on item strengthening (e.g., Pyc & Rawson, 2009) and the expected number of items to be strengthened (e.g., Kornell et al., 2011); previous accounts have provided only fragments of evidence for the overall framework that we present here.

One way to make tests easier or harder is to provide more or fewer retrieval cues. For example, imagine a learner who is encoding a rather difficult set of items: English–Iñupiaq word pairs (e.g., tea–*saiyu*). Rather than have this learner attempt full retrieval of the targets, we could instead provide her with some letters of the target and have her retrieve the rest (e.g., tea–*s__y_*). If this learner plans to test herself multiple times, we could provide her with progressively fewer and fewer cues with each retrieval attempt. That is, we could *scaffold* retrieval demands such that facilitative cues would be gradually removed until our learner could provide target information on her own, without assistance. The three panels in Fig. 1b illustrate a hypothetical progression of cue impoverishment required to maintain an optimal retrieval depth over multiple retrieval attempts (note that the exact shape of the function relating retrieval probability and memory strengthening was chosen for illustrative purposes; we are not making any claims about the form of the relationship other than that it should be an inverted U). Early in practice, our learner may struggle to retrieve any sizable portion of the Iñupiaq target, and so a nearly intact cue will be most beneficial for that item (Fig. 1b, left panel). With additional practice, the probability of retrieval success for the various levels of cuing grows so that more impoverished cues become a more optimal means of enhancing memory (1B, middle panel). Eventually, a full retrieval of the target will be most beneficial (Fig. 1b, right panel).

Fiechter and Benjamin (2018) reported a series of experiments that evaluated a technique designed to accommodate a scenario like the one illustrated in Fig. 1, which they called *diminishing-cues retrieval practice* (DCRP; see also Finley, Benjamin, Hays, Bjork, & Kornell, 2011). In DCRP, learners are initially exposed to a complete cue–target pair, and, over subsequent practice rounds, letters are randomly omitted from the target, one at a time, until the learner must finally make a

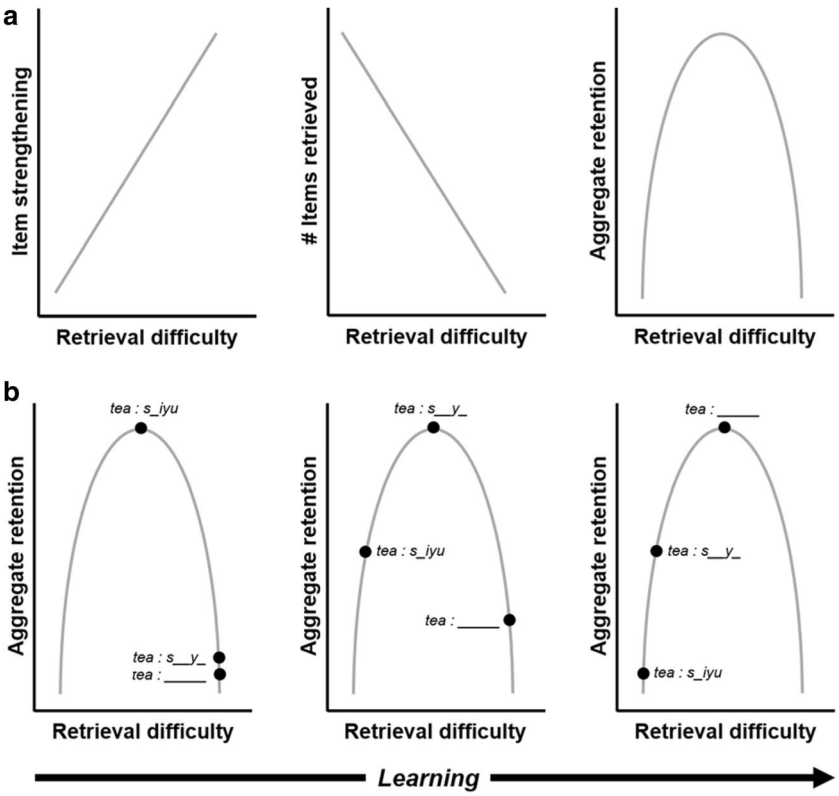


**Fig. 1** A schematic demonstrating the relationship between retrieval difficulty and subsequent memory strengthening. More difficult retrieval results in greater strengthening, but for fewer items, resulting in an inverted-U relationship between test difficulty and aggregate retention (**a**). Retrieval should therefore be neither too difficult nor too easy. With more learning, fewer cues are required to maintain an optimal level of retrieval difficulty (**b**)

full retrieval of the target word with no assistance. Fiechter and Benjamin (2018) found that DCRP was more beneficial than standard testing in the absence of feedback, and it was just as effective as testing when feedback was provided. The DCRP technique therefore successfully extended the benefits of testing to learning scenarios in which testing is not normally effective.

DCRP was designed to accommodate a progression much like the one illustrated in Fig. 1. However, the amount of learning for a particular item on any one practice trial is likely to vary from that rigid series of events: Some items may be well learned on the very first practice attempt (as illustrated in the right panel); other items may prove to be subjectively quite difficult over the entirety of practice (as illustrated in the left panel). With an eye toward this issue, we evaluate here a more flexible approach to scaffolding called *adaptive-cues retrieval practice* (ACRP), which tailors retrieval demands to the current level of mastery for each item. ACRP generates an optimally difficult retrieval attempt by first having a learner attempt a retrieval in the absence of any cuing and then provide increasingly informative cues until the learner would be able to retrieve the correct response. That is, it provides just enough assistance to learners in order for them to retrieve the correct response.

Similar adaptive cuing has been shown to be a beneficial intervention in amnesic patient populations (e.g., Glisky, Schacter, & Tulving, 1986) and with healthy young adults (Finn & Metcalfe, 2010; Hays, 2009). Finn and Metcalfe (2010) found that adaptive cuing benefitted learners' retention of trivia question responses. However, they analyzed only those items that learners initially incorrectly retrieved, and so their data do not speak to the overall benefits of adaptive cuing relative to standard retrieval practice, including those items that are initially correctly recalled. Hays (2009) found that adaptive cuing benefitted retention as much as standard retrieval practice with corrective feedback. However, his adaptive cuing condition did not itself include feedback and so may have underestimated the benefits of that schedule.

The present experiments provide a more comprehensive assessment of the benefits of adaptive cuing by (1) assessing benefits for all items, regardless of whether they were correctly retrieved during practice, and (2) ensuring that corrective feedback was absent or present across all practice conditions being compared. In our experiments that did not include feedback, we predicted that ACRP would more frequently provide learners with an optimal level of cuing (i.e., it will cue learners to a degree that corresponds to the peak of the function in Fig. 1) relative to standard testing and to DCRP, and that this optimal cuing would result in greater long-term retention. In our experiments that provided feedback, we predicted that ACRP and DCRP would be at least as effective as standard retrieval practice (under the assumption that feedback neutralizes differences in retrieval success; Kornell, Klein, & Rawson, 2015), and, to the extent that enhanced retrieval success promotes retention in the presence of feedback (e.g., Metcalfe & Finn, 2010), ACRP would also be expected to outperform standard testing.

## General method

### Subjects

Participants in all our experiments were recruited from Amazon's Mechanical Turk service and were paid for their participation. For each experiment, we initially recruited 80 participants to complete the first part of the experiment and only analyzed the data from subjects who returned to complete the test phase. (We did not use any exclusion rules; data from all participants who completed both phases were analyzed.) This recruiting policy was motivated by a power analysis that was conducted with the aim of replicating the smallest effect size observed in Finn and Metcalfe's (2010) experiments with 80% power. This power analysis suggested a sample size of 44, but because of methodological differences between their experiments and ours, we erred on the side of caution and elected to instead recruit 80 people and then analyze the data from those that ultimately completed both parts (anticipating that at least 50 participants would return for both parts based on previous experiments that we have conducted on Mechanical Turk). The median age for all participants who completed the first day was 32 years; the age range was 19 to 73 years (only 23 of these 480 participants were 55 or older; any differential effects between younger and older adults were therefore expected to have played a minimal role in our findings). The ages of our participants in each experiment are displayed in Table S4 in the online Supplementary Materials.

### Design

All experiments consisted of a single manipulation of practice schedule, resulting in either a two-level or three-level within-subjects design (practice schedule: ACRP, DCRP, standard retrieval practice, or restudy), depending on how many practice schedules were being compared.

### Materials

Our stimuli consisted of 12 English–Iñupiaq word pairs (e.g., tea–*saiyu*) from Finley et al. (2011). Every target word was five letters long.

**Table 1.** Summary of experiments in terms of provision of feedback, sample size, and final test performance

| Experiment | Feedback | *N* | ACRP | DCRP | RP | S |
|---|---|---|---|---|---|---|
| 1a | No | 62 | 0.42 (0.33)[b] | – | 0.15 (0.25) | 0.23 (0.33) |
| 2a | No | 58 | 0.38 (0.36)[b] | 0.32 (0.31)[b] | – | – |
| 3a | No | 60 | 0.52 (0.34)[b] | 0.37 (0.33) | – | – |
| 1b | Yes | 69 | 0.52 (0.36)[b] | – | 0.53 (0.34)[b] | 0.31 (0.31) |
| 2b | Yes | 57 | 0.44 (0.36)[b] | 0.44 (0.36)[b] | – | – |
| 3b | Yes | 63 | 0.62 (0.34)[b] | 0.56 (0.36)[b] | 0.57 (0.37)[b] | – |

*Note*. The four rightmost columns indicate proportion correct in the applicable conditions (standard deviations are in parentheses). ACRP = adaptive-cues retrieval practice; DCRP = diminishing-cues retrieval practice; RP = standard retrieval practice; S = restudy

[b] Best performance in a given experiment. Values are considered tied if they are unconvincingly different from one another (i.e., $BF_{10}$ <3)

## Procedure

Based on effect sizes reported by Rowland (2014), we conducted our experiments under two conditions: The first set, which was not expected to yield a testing effect, used items whose initial retrievability was less than 50%, and withheld feedback; the second, which was expected to yield a strong testing effect, again used items with low initial retrievability, but now we provided feedback to our learners. A summary of the features of each experiment is presented in Table 1.

**Study phase** Participants initially studied each word pair for 4 seconds. They cycled through the list of items three times; each cycle followed the same random order. After the study phase, participants performed a go/no-go distractor task for 1 minute before moving on to the practice phase. For this task, letters were presented for 750 milliseconds at random intervals, and participants were to avoid pressing the space bar if the presented letter was an *X*.

**Practice phase** Practice consisted of a within-subjects manipulation of practice type; depending on whether an experiment compared two or three practice conditions, six or four items each were randomly assigned to each condition. Four practice conditions were compared, in various combinations, across our three series of experiments. In the restudy condition, subjects were presented with the complete English–Iñupiaq pair and asked to type in the target. In the retrieval practice condition, subjects were presented with the English cue paired with five blank spaces and asked to provide the Iñupiaq target. The ACRP condition in Experiments 1a through 2b was implemented as follows: Subjects were shown the English target and five blank spaces; they were asked to either provide the Iñupiaq target or else press the space bar to request a letter. Letter presentation was randomly ordered. Letter requests could be made only after 2 seconds had elapsed since the last request. Unlike Hays (2009) and Finn and Metcalfe (2010), learners were not required to submit a correct response to proceed to the next practice trial. Rather, the program would accept any submission regardless of its correctness. In Experiments 3a and 3b, we revised the ACRP condition so that learners had to provide a response before they would receive an additional letter. They would then be asked to provide another response, and, in the case of an incorrect response, would be given another letter. This routine continued until the correct response was provided (responses had to be exactly correct; we did not use lenient scoring) or until all five letters of the target were shown. In the DCRP condition, learners were initially shown a complete cue–target pair. Letters were randomly omitted, one at a time, from the target with each subsequent practice round until no letters remained by the final round of practice.

Practice consisted of six rounds;[1] each round involved cycling through the items—and, consequently, their respective practice conditions—in a newly randomized order, with the constraint that items presented in the first and second halves of the first round remained in their respective halves throughout the rest of practice. All practice trials were self-paced. For experiments in which feedback was provided, learners saw the complete English–Iñupiaq pair for 4 seconds after submitting a response. Feedback was withheld in our "a" experiments and provided in our "b" experiments. The first session ended following completion of the practice phase.

**Final test** Subjects could complete the test session 12 to 36 hours after completing the first part of the experiment (the average retention interval was 19.25 hours [*SD* = 5.14]). During this session, participants were given the English cues and asked to provide the Iñupiaq target. All test trials were self-paced. Once they had completed this final test, participants were thanked and given access to a digital debriefing form.

[1] We initially conducted two experiments that were identical to Experiments 1a and 1b, except we provided only one practice round. Performance between conditions did not differ; we attributed these null findings to a floor effect resulting from insufficient practice (overall, only 19% of items were correctly recalled at test). All subsequent experiments therefore included six rounds of practice.

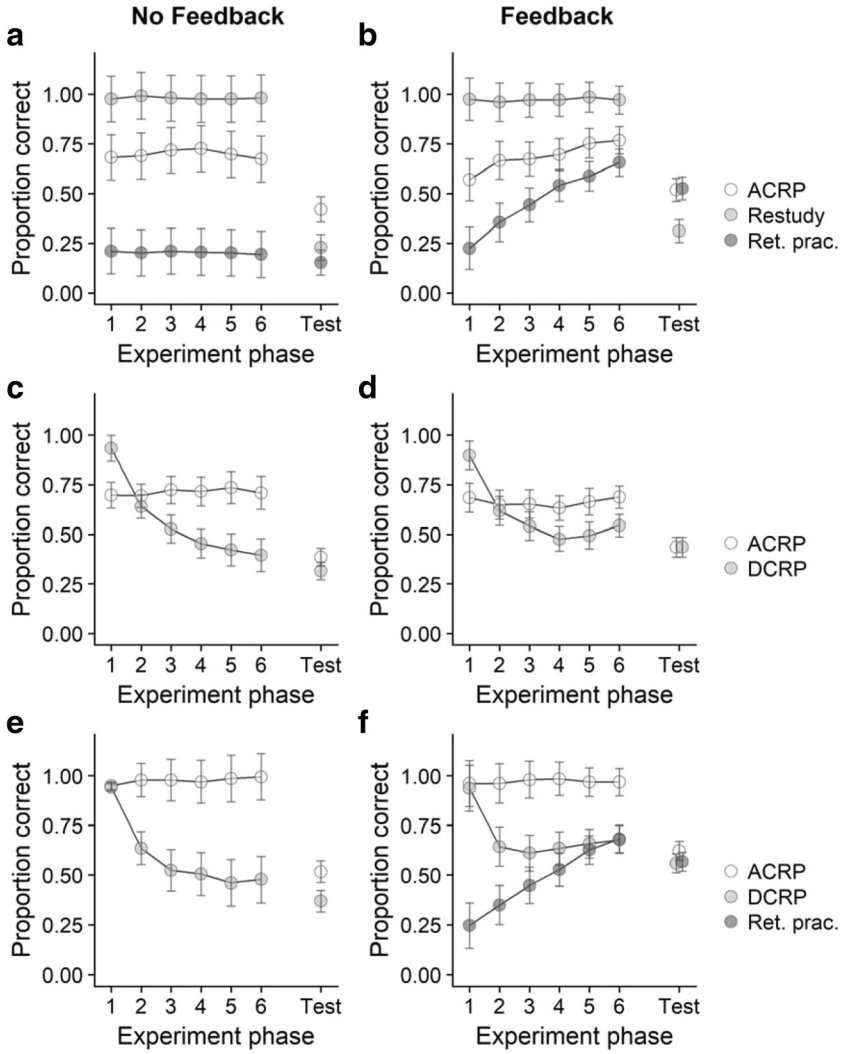


**Fig. 2** Accuracy in each practice round (indicated by numbers 1–6 along the abscissa) and at test as a function of practice condition in Experiments 1a and 1b (**a** and **b**), 2a and 2b (**c** and **d**), and 3a and 3b (**e** and **f**). Height of the error bars indicates within-subject 95% confidence intervals (Morey, 2008)

## Experiments 1a and 1b

Our first two experiments evaluated the benefits of ACRP relative to standard testing and a restudy control condition, with and without feedback.

### Results

All data are available at the Open Science Framework at https://osf.io/9jm54/. We analyzed our data with Bayesian *t* tests (Rouder, Speckman, Sun, Morey, & Iverson, 2009). Unlike null hypothesis significance testing, which only speaks to the probability of data under the assumption that the null hypothesis is true, Bayesian analyses allow for evaluations of evidence in favor of both the null and alternative hypotheses. Specifically, our analyses evaluated the likelihood of a point null hypothesis (i.e., a Cohen's *d* of zero) to that of a Jeffrey–Zellner–Siow alternative prior (i.e., a Cauchy distribution of *d*s; Rouder et al., 2009). Bayesian analyses have the additional benefit of requiring no adjustments for multiple comparisons, as long as subject-level variance is accounted for (Gelman, Hill, & Yajima, 2012).

We report Bayes factors, which are ratios of evidence in favor of these null and alternative hypotheses. All Bayes factors are reported in terms of evidence favoring the alternative hypothesis: values greater than one indicate evidence favoring the alternative, and values less than one indicate evidence favoring the null. Following recommendations by Jeffreys (1961), we interpret Bayes factors greater than 3 and less than 0.33 as the minimum criteria for evidence in favor of the alternative and null hypotheses, respectively. Values not meeting these criteria are considered unconvincing in their support of the presence or absence of an effect. For all comparisons involving the restudy condition, we used one-tailed tests

**Table 2.** Total practice time per item, in seconds, in each condition of each experiment

| Experiment | ACRP | DCRP | RP | S |
|---|---|---|---|---|
| 1A | 68.90 (30.58)[d] | – | 45.14 (22.59)[d] | 24.34 (9.33)[d] |
| 2A | 71.03 (32.88)[d] | 40.90 (20.50)[d] | – | – |
| 3A | 94.57 (47.59)[d] | 37.54 (18.78)[d] | – | – |
| 1B | 71.95 (44.13)[d] | – | 49.68 (23.42)[d] | 37.88 (41.73) |
| 2B | 63.37 (23.29)[d] | 42.37 (16.49)[d] | – | – |
| 3B | 95.81 (46.87)[d] | 38.92 (16.90)[d] | 43.93 (18.66)[d] | – |

*Note.* Values are the means of participant median times. Standard deviations are indicated in parentheses. ACRP = adaptive-cues retrieval practice; DCRP = diminishing-cues retrieval practice; RP = standard retrieval practice; S = restudy

[d] Values were convincingly different (i.e., $BF_{10}$ >3) from practice times in the other condition(s)

because we were only interested in whether our other conditions were superior to restudy or not.

**Practice and final test performance** Practice and retention data from Experiment 1a are presented in Fig. 2a (analyses of practice performance are included in the online Supplementary Materials). Final test performance in the ACRP condition ($M$ = 0.42, $SD$ = 0.33) was superior to performance in the restudy condition ($M$ = 0.23, $SD$ = 0.33, $BF_{10}$ = 1,940.67, $d$ = 0.83) and to performance in the retrieval practice condition ($M$ = 0.15, $SD$ = 0.25, $BF_{10} = 8.78 \times 10^6$, $d$ =1.31).[2] As planned, performance in the retrieval practice condition was not superior to the restudy condition ($BF_{10}$ = 0.05).

Practice and retention data from Experiment 1b are presented in Fig. 2b. Performance in the ACRP condition ($M$ = 0.52, $SD$ = 0.36) was once again superior to performance in the restudy condition ($M$ = 0.31, $SD$ = 0.31, $BF_{10} = 4.75 \times 10^4$, $d$ = 0.94), but was now approximately the same as performance in the retrieval practice condition ($M$ = 0.53, $SD$ = 0.34, $BF_{10}$ = 0.14). As planned, we also observed a strong testing effect ($BF_{10} = 7.40 \times 10^5$, $d$ = 1.06).

**Practice time** The total practice time per item in each condition is reported in Table 2. In Experiment 1a, items in the ACRP condition were practiced for longer than items in both the restudy condition ($BF_{10} = 2.18 \times 10^{18}$, $d$ = 2.52) and the retrieval practice condition ($BF_{10} = 1.27 \times 10^6$, $d$ = 1.22). Additionally, tested items were practiced for longer than restudied items ($BF_{10} = 8.51 \times 10^9$, $d$ = 1.97).

In Experiment 1b, items in the ACRP condition were once again practiced for longer than items in either the retrieval practice ($BF_{10}$ = 593.65, $d$ = 0.80) or restudy conditions ($BF_{10} = 1.54 \times 10^9$, $d$ = 1.42). The difference in practice time between tested and restudied items yielded ambiguous evidence ($BF_{10}$ = 1.39). These findings raise the possibility that the performance advantage for ACRP may have arisen merely from the fact that learners were spending more time on items in that condition. However, we reanalyzed our final test performance data with practice time as a covariate and our results remained unchanged.[3]

### Discussion

We observed an advantage for ACRP over the other two conditions when feedback was not provided, and similar performance between ACRP and retrieval practice when feedback was provided. The DCRP technique described earlier (Fiechter & Benjamin, 2018) was also superior to standard retrieval practice without feedback and equally effective when feedback was provided. Because the benefits of ACRP yielded similar data patterns to those observed with DCRP, we next compared ACRP and DCRP directly, with and without feedback, to see if they differentially enhanced memory.

## Experiments 2a and 2b

Our next two experiments directly compared the benefits of ACRP and DCRP to one another, with and without feedback.

### Results

**Practice and final test performance** Practice and retention data from Experiments 2a and 2b are presented in Fig. 2c and d. In Experiment 2a, final test performance in the ACRP condition ($M$ = 0.38, $SD$ = 0.36) was not

[2] For all no-feedback experiments (1a, 2a, and 3a), we also observed benefits of ACRP after removing those items that received all five letters of the target at any point in practice. The robustness of this effect to this additional analysis reveals that the benefits don't arise from advantages accruing to subjects who use the procedure to elicit a "feedback" (i.e., the correct answer) without genuine attempts to retrieve the terms initially.

[3] We included time on task as a covariate in regression analyses of all our experiments, and all results were unchanged. These analyses are included in the Supplementary Materials.

convincingly different from DCRP ($M$ = 0.32, $SD$ = 0.31, $BF_{10}$ = 1.40). Results from Experiment 2b were similar, as ACRP performance ($M$ = 0.44, $SD$ = 0.36) was once again not different from DCRP ($M$ = 0.44, $SD$ = 0.36; $BF_{10}$ = 0.14).

**Practice time** The total practice time per item in each condition is reported in Table 2. In Experiment 2a, items were practiced for longer in the ACRP condition relative to the DCRP condition ($BF_{10} = 1.91 \times 10^{12}$, $d$ = 1.91). We obtained the same result in Experiment 2b ($BF_{10} = 4.18 \times 10^{8}$, $d$ = 1.62).

### Discussion

With and without feedback, we found no convincing performance differences between ACRP and DCRP, and much shorter practice times for items in the DCRP condition. However, as noted at the outset, our implementation of ACRP was slightly different from that used by Hays (2009) and Finn and Metcalfe (2010). In their experiments, subjects had to attempt a response before they received assistance; in our experiments, participants requested letters by pressing the space bar. Our initial ACRP method therefore placed the onus on participants to judge when they were ready to retrieve the target words. Unfortunately, learners often lack the metacognitive sophistication to successfully self-regulate their learning (see Fiechter, Benjamin, & Unsworth, 2016) as we had allowed them to in our first four experiments. With an eye toward this potential limitation, we revised our ACRP method so that learners now had to provide a response prior to provision of a letter. This revision ensured that participants would be retrieving at their optimal level of cuing (i.e., the level at which they would first be able to successfully produce the target) because they were now forced to provide a (possibly correct) response even when they were uncertain about the accuracy of it. In contrast, our initial implementation of ACRP allowed learners to delay retrieval until they were certain of the correct response, which may have deprived them of a deeper retrieval.

## Experiments 3a and 3b

Our final two experiments evaluated the benefits of the revised ACRP technique against DCRP, without feedback (3a), and against DCRP and standard retrieval practice, with feedback (3b). We included the retrieval practice condition in 3b in order to assess whether the revised ACRP condition might benefit memory more than testing with feedback, as Finn and Metcalfe (2010) have shown with a similar adaptive cuing technique.

### Results

**Practice and final test performance** Practice and retention data from Experiments 3a and 3b are presented in Fig. 2e and f. In Experiment 3a, final test performance in the ACRP condition ($M$ = 0.52, $SD$ = 0.34) was now superior to performance in the DCRP condition ($M$ = 0.37, $SD$ = 0.33; $BF_{10}$ = 341.82, $d$ = 0.79).[4]

In Experiment 3b, performance in the ACRP condition ($M$ = 0.62, $SD$ = 0.34) was not convincingly different from the DCRP condition ($M$ = 0.56, $SD$ = 0.36, $BF_{10}$ = 1.04) and retrieval practice condition ($M$ = 0.57, $SD$ = 0.37, $BF_{10}$ = 0.48). Performance in the DCRP condition was approximately the same as the retrieval practice condition ($BF_{10}$ = 0.14).

**Practice time** The total practice time per item in each condition is reported in Table 2. In Experiment 3a, items were practiced for longer in the ACRP condition relative to the DCRP condition ($BF_{10} = 8.08 \times 10^{9}$, $d$ = 1.69). In Experiment 3b, ACRP items were once again practiced for longer than DCRP items ($BF_{10} = 7.80 \times 10^{12}$, $d$ = 2.18) and tested items ($BF_{10} = 7.44 \times 10^{10}$, $d$ = 1.85). Importantly, DCRP items were practiced for less time than tested items ($BF_{10}$ = 5.62, $d$ = 0.51).

### Discussion

The revised version of ACRP, in which learners had to provide a response prior to receiving additional cuing, was superior to DCRP when learners were without feedback. However, the ACRP technique was even more time-consuming than in previous experiments. When feedback was provided, we found no convincing differences in performance between ACRP, DCRP, and retrieval practice. Critically, DCRP required less total practice time per item than did either ACRP or standard retrieval practice. Thus, although our data suggest that it is indeed beneficial to differentially allocate retrieval demands based on an item's current memory strength, they also suggest that substantial time is required to ensure optimal retrieval difficulty (at least as we have implemented the ACRP technique here). When feedback is present, DCRP may offer the best combination of encoding efficiency and mnemonic benefits.

## General discussion

To recap, our initial version of ACRP—which required learners to press the space bar to request assistance—produced similar benefits to DCRP: Under conditions in

[4] The advantage of ACRP over DCRP was evident even after including number of item-level retrieval attempts as a covariate in a logistic-regression analysis. This analysis is presented in the online Supplementary Materials.

which testing was not effective, ACRP was effective; under conditions where the benefits of testing were robust, ACRP enhanced memory to approximately the same extent as both testing and DCRP. A direct comparison of ACRP and DCRP in our next two experiments suggested that these conditions enhanced memory to a similar degree, with DCRP yielding much shorter learning times than ACRP. In the final two experiments, we revised ACRP so that learners had to provide a full response before they could receive additional assistance. This revised ACRP enhanced memory even more than DCRP without feedback; ACRP, DCRP, and standard retrieval practice all enhanced memory to a similar degree with provision of feedback, while DCRP once again yielded the shortest study times.

Relative to DCRP, ACRP enhanced retention when feedback was not present. This result aligns with the assertion that retrieval is most beneficial when retrieval is neither too easy nor too difficult (Benjamin & Tullis, 2010). We therefore attribute the superiority of ACRP to the more optimal levels of cuing that it achieves (that is, we believe that cuing in ACRP more closely corresponds to the apex of the function in Fig. 1b). When feedback was present, ACRP, DCRP, and standard retrieval practice all benefited memory to approximately the same degree. That the enhanced retrieval success of ACRP and DCRP—relative to standard testing—was not advantageous with provision of feedback supports the argument that exposure to correct information, and not retrieval success specifically, drives the benefits of testing with feedback (Kornell et al., 2015).

DCRP required less time during practice than did ACRP or standard testing, suggesting that progressively diminished cuing offers an efficiency advantage without trading off mnemonic benefits. In fact, it may be the case that our data underestimate the superior efficiency of the DCRP schedule relative to standard retrieval practice with feedback: The greater retrieval success of DCRP during practice suggests that one would require provision of feedback much less frequently than with standard testing, thereby saving even more time. Participants in Experiment 3b were required to view a complete pair for four seconds after every retrieval attempt and so we were not able to assess potential time savings stemming from differential time spent processing feedback.

ACRP could also be adjusted to make that technique more efficient. For example, Glisky et al. (1986) implemented an adaptive cuing technique that omitted or added letters to a target across trials (rather than within trial, as we did) depending on whether learners had successfully retrieved, or not, the target on the previous practice trial. This approach to adaptive cuing would almost certainly require less practice time than did our implementation of ACRP, but whether the mnemonic benefits of ACRP would be preserved is an open question. Future work should continue to investigate means of implementing retrieval practice that are more efficient and, if possible, enhance memory to a greater degree than does standard testing with feedback.

Future work should also evaluate techniques such as DCRP and ACRP with more complex materials. For example, using materials such as term-definition pairs, key terms or phrases of a definition could be omitted or added to generate optimally difficult retrieval attempts. Alternative methods of scaffolding retrieval should also be evaluated. Past work has sought to optimize test difficulty by providing helpful cues (as we have done) and by manipulating the intervals between tests to make them appropriately difficult (e.g., Landauer & Bjork, 1978). However, a method of scaffolded retrieval that has not been evaluated is updating test formats so that they become progressively more difficult. For example, learners' knowledge could initially be assessed with a multiple-choice test, and then fill-in-the-blank responses, and finally a free-response format with no external prompting.

The question of how to best implement testing has important ramifications for any learning situation, and especially educational practice. We have presented evidence suggesting that (1) adaptive learning schedules are effective but may require additional fine-tuning to make them more efficient, and (2) scaffolded but nonadaptive retrieval is faster and just as beneficial as self-testing with feedback. As testing and technology are increasingly relied upon in the classroom, the testing regimens evaluated here offer a means of leveraging both in the service of enhancing educational practice.

**Acknowledgements** We thank the Human Memory an Cognition Laboratory at Illinois for help with design and analysis.

## References

Abbot, E. E. (1909). On the analysis of the factor of recall in the learning process. *Psychological Monographs*, *11*, 159–177.

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology*, *61*, 228–247.

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

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

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*, 1095–1102.

Fiechter, J. L., & Benjamin, A. S. (2018). Diminishing-cues retrieval practice: A memory-enhancing technique that works when regular testing doesn't. *Psychonomic Bulletin & Review, 25*(5), 1868–1876. https://doi.org/10.3758/s13423-017-1366-9.

Fiechter, J. L., Benjamin, A. S., & Unsworth, N. (2016). The metacognitive foundations of effective remembering. In J. Dunlosky & S. K. Tauber (Eds.), Oxford handbook of metamemory (pp. 301–324). New York: Oxford University Press.

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.

Finn, B. & Metcalfe, J. (2010). Scaffolding feedback to maximize long-term error correction. *Memory & Cognition*, *38*, 951–961.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*, 189–211.

Glisky, E. L., Schacter, D. L., & Tulving, E. (1986). Learning and retention of computer-related vocabulary in memory-impaired patients: Method of vanishing cues. *Journal of Clinical and Experimental Neuropsychology*, *8*, 292–312.

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

Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: The role of retrievability. *The Quarterly Journal of Experimental Psychology*, *65*, 962–975.

Jeffreys, H. (1961). Theory of probability (3rd). Oxford: Oxford University Press.

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.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial in Quantitative Methods for Psychology*, *4*, 61–64.

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.

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.

## Open practices statement

Data from all experiments and an accompanying R script are available at the Open Science Framework at https://osf.io/9jm54/. None of the experiments was preregistered.