Journal of Memory and Language 60 (2009) 437–447

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory?

Mary A. Pyc *, Katherine A. Rawson

Kent State University, Department of Psychology, P.O. Box 5190, Kent, OH 44242-0001, United States

ARTICLE INFO

Article history:
Received 19 August 2008
Revision received 29 December 2008
Available online 27 February 2009

Keywords:
Desirable difficulties
Retrieval practice
Memory
Testing effects

ABSTRACT

Although substantial research has demonstrated the benefits of retrieval practice for promoting memory, very few studies have tested theoretical accounts of this effect. Across two experiments, we tested a hypothesis that follows from the desirable difficulty framework [Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe, A. Shimamura, (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press], the *retrieval effort hypothesis*, which states that difficult but successful retrievals are better for memory than easier successful retrievals. To test the hypothesis, we set up conditions under which retrieval during practice was successful but differentially difficult. Interstimulus interval (ISI) and criterion level (number of times items were required to be correctly retrieved) were manipulated to vary the difficulty of retrieval. In support of the retrieval effort hypothesis, results indicated that as the difficulty of retrieval during practice increased, final test performance increased. Longer versus shorter ISIs led to more difficulty retrieving items, but higher levels of final test performance. Additionally, as criterion level increased, retrieval was less difficult, and diminishing returns for final test performance were observed.

Introduction

Interest in the benefits of retrieval practice for subsequent memory has increased dramatically in recent years due to important implications for student learning and scholarship. A wealth of research has indicated that retrieval practice can be used not only as a means to assess memory but as an effective means to improve memory (for a recent review see Roediger & Karpicke, 2006).

Although demonstrations that retrieval practice is beneficial for promoting memory are increasingly numerous, the extant literature is largely empirical rather than theoretical at this point. This is not to say that theoretical frameworks relevant to explaining effects of retrieval practice do not exist, and some recent work has reviewed existing findings in light of these accounts (e.g., Bjork, 1994; Carpenter, Pashler, & Vul, 2006; Roediger & Karpicke, 2006). However, very few studies have been designed to directly test a priori predictions of proposed theories (Carpenter & DeLosh, 2006; Glover, 1989; McDaniel & Masson, 1985). Put differently, retrieval practice effects are well documented, but the factors that underlie the effects are less well established. Accordingly, the goal of the current research was to provide a theoretical advance to the extant literature. To foreshadow, we first describe the general theoretical framework motivating the current work. We then introduce a specific hypothesis that follows from this framework and two experiments designed to directly test a priori predictions from the hypothesis.

The present work was motivated by the *desirable difficulty* framework (e.g., Bjork, 1994, 1999). The general principles of the framework specify that within any learning task or domain, difficult but successful processing will be better for memory than difficult but unsuccessful processing, a relatively intuitive claim. The more provocative claim is that successful but difficult processing will be better for memory than successful but easier processing.

* Corresponding author.
E-mail address: mpyc@kent.edu (M.A. Pyc).

0749-596X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2009.01.004

Of course, applying this general framework to a particular task domain requires a specific instantiation of its claims that are appropriate to the learning task of interest. Although frameworks cannot be tested directly, specific hypotheses instantiated to apply the basic principles of the framework to a particular learning task can be tested. Accordingly, to apply the desirable difficulty framework to retrieval practice, the specific instantiation of these general principles tested here is the *retrieval effort hypothesis*. The basic claim of the retrieval effort hypothesis is that not all successful retrievals are created equal: given that retrieval is successful, more difficult retrievals are better for memory than less difficult retrievals.

Thus, two conditions must be satisfied to directly test the retrieval effort hypothesis. First, retrieval during practice must be successful (hereafter we will simply use the term *retrieval* to refer to retrieval during practice). To satisfy this condition in the current work, items were practiced until they were correctly retrieved a predetermined number of times. Second, difficulty of retrieval must vary. To satisfy this condition, we manipulated two variables. The first variable was *interstimulus interval* (ISI, defined here as the number of items between each next practice trial with any given item), and the second variable was *criterion level* (the number of times items were required to be correctly recalled before dropping from practice). Each of these manipulations is based on an assumption about the relationship between the manipulated factor and retrieval difficulty. Below, we discuss each assumption in turn, followed by the prediction from the retrieval effort hypothesis that rests on that assumption.

The first assumption is that correct retrieval of items is more difficult after a longer ISI than after a shorter ISI (hereafter referred to as the *ISI assumption*). Evidence supporting this assumption comes from recent work by Karpicke and Roediger (2007a), in which response latencies for correct retrievals during practice were shorter for items correctly retrieved after an ISI of zero items (i.e., massed trials) compared to items retrieved after a longer ISI (either one or five intervening items).

Based on the ISI assumption, the retrieval effort hypothesis predicts that final test performance will be greater for items correctly retrieved after a longer ISI than items correctly retrieved after a shorter ISI (hereafter referred to as the *ISI prediction*). Regarding previous findings bearing on the ISI prediction, many studies have manipulated ISI. However, previous research has manipulated ISI between a fixed number of practice trials, which would obviously then involve a mixture of trials in which items were correctly retrieved and trials in which items were not correctly retrieved. In contrast, no previous study has examined the effects of ISI between *correct retrievals*, which provides a stronger test of the retrieval effort hypothesis.

The second assumption is that as the number of times an item is correctly retrieved (i.e., criterion) increases, the difficulty of each next correct retrieval will decrease (hereafter referred to as the *criterion assumption*). Evidence supporting this assumption also comes from recent work by Karpicke and Roediger (2007a) reporting response latencies for practice test trials in which the item was correctly retrieved. Results indicated that as the number of correct retrievals increased (ranging from one to three), response latencies decreased.

Based on the criterion assumption, the retrieval effort hypothesis predicts that as the number of times items are correctly retrieved increases, the incremental benefit to final test performance will decrease; that is, a curvilinear relationship between number of correct retrievals and final test performance is predicted (hereafter referred to as the *criterion prediction*). Note that this is not to imply that more correct retrievals will not enhance final test performance; a reasonable expectation is that more correct retrievals will lead to higher levels of final test performance than fewer. Rather, the retrieval effort hypothesis predicts a greater increase in final test performance from earlier versus later correct retrievals, because difficulty of retrieval is greater earlier in learning compared to later in learning.

Regarding previous findings bearing on the criterion prediction, research on overlearning has typically manipulated the number of trials or amount of practice time rather than the number of correct retrievals during practice (e.g., Kratochwill, Demuth, & Conzemius, 1977; Rohrer, Taylor, Pashler, Wixted, & Cepeda, 2005). Only one previous study has manipulated the number of correct retrievals during practice (Nelson, Leonesio, Shimamura, Landwehr, & Narens, 1982). Using paired associates (e.g., 48-dollar), Nelson et al. (1982) required participants to retrieve items one, two, or four times. On the final test four weeks later, performance increased as the number of correct retrievals increased. However, the fairly limited range of criterion levels (1, 2, or 4 correct retrievals) makes it difficult to determine whether the relationship between final test performance and the number of times items were correctly retrieved is linear or curvilinear. Karpicke and Roediger (2007b) have examined final test performance as a function of the number of times items were correctly retrieved with a larger range of criterion levels. Their participants learned word lists using conditions in which items could be correctly recalled up to 15 times. Of interest here, results indicated a curvilinear relationship between final test performance and the number of times items were correctly retrieved. However, because the primary interest in their study was not in evaluating the effect of increasing the number of correct retrievals, these analyses were conducted post hoc (see also Pyc, Rawson, & A., submitted for publication). Therefore, these results are difficult to interpret because items were not assigned to criterion level. Items correctly retrieved more times were likely the easier items, and the results may have been due in part to item difficulty effects. Accordingly, to extend beyond previous research, we manipulated the number of times items were required to be correctly retrieved before dropping from test-restudy practice, with a wider range of criterion levels. Thus, we were able to evaluate the relationship between the number of times an item is correctly retrieved and subsequent memory performance without concerns for item difficulty effects.

In sum, Experiments 1 and 2 were designed to directly test predictions from the retrieval effort hypothesis. We created conditions in which retrieval would be successful

(items were required to be learned to criterion) but differentially difficult (manipulations of ISI and criterion level). Experiment 2 extended beyond Experiment 1 by including a latency measure to evaluate the ISI and criterion assumptions, which are the bases for the predictions of the retrieval effort hypothesis.

Experiment 1

Method

Participants and design

One hundred twenty-nine participants enrolled in Introductory Psychology at Kent State University participated in return for course credit. ISI (short versus long) was a between-participant manipulation. Criterion level (1, 3, 5, 6, 7, 8, or 10 correct retrievals per item during practice) was a within-participant manipulation. To establish the generality of the retrieval effort hypothesis, we implemented two retention intervals (RI) between practice and final test, 25 min (short RI) or one week (long RI) as a between-participant variable.

Materials

Materials included 70 Swahili-English translation word pairs (previously normed for item difficulty by Nelson & Dunlosky, 1994). Seven items were assigned to each of 10 lists, with an equivalent range of item difficulty within each list. Within each list, one item was randomly assigned to each of the seven criterion levels (randomized anew for each participant). Random assignment of item to criterion level minimizes concerns about the contribution of item difficulty to effects of criterion level (one of the interpretive limitations of previous research).

Procedure

All task instructions and items were presented via computer. All items received both an initial study trial and test-restudy practice trials. On initial study trials, items received a 10 s presentation of both the cue and target. On practice trials, participants were presented with a cue and had 8 s to recall the target (by typing the target in a field provided on the computer screen). If participants recalled items before 8 s had elapsed, they could press a "done" key to advance. If items were correctly retrieved, there was no restudy opportunity for the item on that test trial.[1] Incorrectly retrieved items received a 4 s restudy opportunity before continuing on to the next to-be-learned item. Participants were not explicitly given feedback about the correctness of each response, but they were informed that only items that were incorrectly retrieved would receive a restudy trial. The computer recorded the number of times each item was correctly retrieved, and items continued to be practiced until they reached their assigned criterion level of performance (1, 3, 5, 6, 7, 8, or 10 correct retrievals), after which point they were dropped from further practice. Participants were aware at the outset that they would be tested on items until they reached an "acceptable level of performance", but were not specifically aware of the number of times individual items were to be correctly retrieved.

In the short ISI group, each of the seven items from the first list was presented for an initial study trial. These items were then presented for test-restudy practice trials (i.e., an ISI of six items, approximately 1 min between each practice trial with a given item) until each item reached its assigned criterion level of performance. When an item reached criterion, it was dropped from test-restudy practice. After all items on the first list reached their assigned criterion level, items from the second list were presented for an initial study trial and then test-restudy trials, and so on until all items in each of the 10 lists reached their criterion level of performance. Order of list presentation was counterbalanced across participants.

In the long ISI group, five of the lists were assigned to the first block of study, and the other five lists were assigned to the second block (with assignment of list to block counterbalanced across participants). All 35 items in the first block were presented for an initial study trial. These items were then presented for test-restudy trials (i.e., an ISI of 34 items, approximately 6 min between each practice trial with a given item) until each item reached its assigned criterion level, at which point it was dropped from test-restudy practice. After all items in the first block reached their assigned criterion level, items from the second block were presented for initial study and then received test-restudy practice trials until each item reached criterion. In both ISI groups, participants were given up to 90 min to learn all items to criterion (to keep the overall length of the experiment reasonable).

Following the practice phase, all groups completed a 25-min reading comprehension filler task that was not related to the main experimental task. Upon completion of the filler task, participants in the short RI group completed the self-paced final cued-recall test for all 70 items. The cued-recall test was the same format as test trials during practice, with the exception that no restudy opportunity was provided for incorrectly recalled items. Participants in the long RI group were dismissed and returned one week later to complete their final cued-recall test.

Results and discussion

The mean proportion of items correctly recalled on the final cued-recall test is presented in Fig. 1, as a function of ISI, criterion level, and retention interval. For ease of exposition, the results of a 2 (long versus short ISI) × 2 (short versus long RI) × 7 (criterion level) mixed factor analysis of variance (ANOVA) are reported in Table 1. 95% confidence intervals (CI) around the difference between means (M_d) are reported for paired comparisons of interest below.

According to the ISI prediction of the retrieval effort hypothesis, final test performance will be greater for items that were retrieved after a longer versus shorter ISI. Confirming this prediction, a significant main effect of ISI indi-

[1] Given the number of items used and the number of times items had to be correctly retrieved, we decided to save time where possible to keep the overall length of the practice phase reasonable. We chose to drop restudy for correct items based on previous research suggesting that feedback does not further improve memory for items that are correctly retrieved during practice (Pashler, Cepeda, Wixted, & Rohrer, 2005).

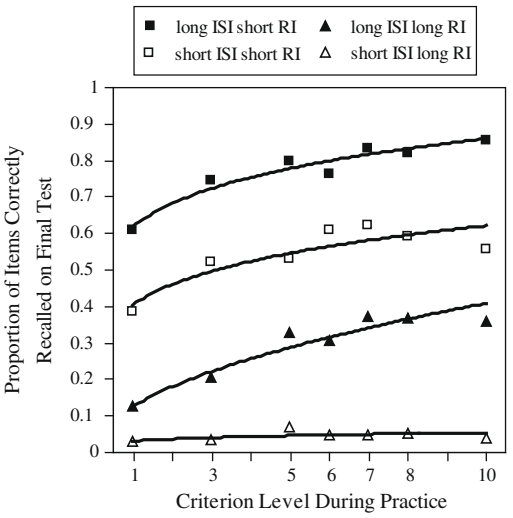

Fig. 1. Mean proportion of items correctly recalled on the final cued-recall test for each group in Experiment 1. Data points represent mean proportion of items correctly recalled at each criterion level. Lines represent the best power function fit for each group across criterion levels.

Table 1
2 × 2 × 7 Mixed ANOVA results for final test performance, Experiment 1.

	df	*F*	*p*
Main effects			
Criterion level	5.28, 654.3[a]	33.92	<.001
Long/short ISI	1, 124	61.74	<.001
Long/short RI	1, 124	254.97	<.001
Interactions			
Criterion level ∗ ISI	5.28, 654.3[a]	5.66	<.001
Criterion level ∗ RI	5.28, 654.3[a]	2.57	.02
Criterion level ∗ ISI ∗ RI	5.28, 654.3[a]	3.87	.001

[a] *Note:* Greenhouse–Geisser corrections are reported to correct for the violation of the assumption of sphericity.

cated that final test performance was greater after longer versus shorter ISIs between correct retrievals (short ISI, $M = .30$, $SE = .02$; long ISI, $M = .54$, $SE = .03$). Presumably, longer ISIs led to more difficult correct retrievals, which in turn led to higher levels of final test performance.

According to the criterion prediction of the retrieval effort hypothesis, as the number of times items are correctly retrieved increases, the incremental benefit to final test performance will decrease. Confirming this prediction, we observed a pattern of diminishing returns from increasing criterion level for final test performance (i.e., less benefit from each next correct retrieval as the number of correct retrievals increased) in three of the four groups. The only exception was the short ISI-long RI group, in which performance was on the floor; final test performance did not significantly differ after correctly retrieving an item once versus ten times during practice. This finding is somewhat surprising and may warrant further investigation in future research; for present purpose, because of this floor effect, further analyses will focus on the other three groups.

Although visual inspection of the general pattern of results supports the criterion prediction, we statistically tested the best function fit to the data to determine whether a linear or curvilinear function provided a better fit. We evaluated function fits at both the group level (means at each criterion level across items and participants within each group) and at the individual level (means at each criterion level across items for each participant) to establish the generality of the effect.

First, we examined linear, power, and exponential function fits for the means at the group level (using SPSS 15.0; see Table 2 for function fits). Both the power and exponential function fits were included because the retrieval effort hypothesis makes no a priori prediction that one would provide a better fit to the data than the other. Rather, we were interested in whether a curvilinear function fit the data better than a linear function. In each of the three groups, the best fitting curvilinear function (the power function) fit the data better than the linear function, accounting for 91% versus 74% of the variance across groups, respectively. The power function fit for the means in each group are shown in Fig. 1.

Second, we computed function fits to the means for each individual to determine whether the pattern observed at the group-level analyses also held at the individual level. We computed the mean across individual R^2 values for each function fit (see Table 2). No differences emerged as a function of group, so we collapsed across the three groups for individual-level analyses. Not surprisingly, the individual-level data were noisier than the group-level data, so the individual R^2 values were lower overall. Nonetheless, the same basic pattern obtained. A power function provided a better fit than the linear function ($M_d = .07$, CI = .03, .11). Taken together, these analyses indicate that the function that best fits the data is curvilinear, which supports the criterion prediction from the retrieval effort hypothesis.

Experiment 2

Results from Experiment 1 supported both predictions of the retrieval effort hypothesis: higher levels of final test performance were observed after longer versus shorter ISIs between correct retrievals. Additionally, the relationship between criterion level and final test performance was curvilinear rather than linear. These results indicate that conditions under which retrieval is successful but more difficult produce greater benefits to memory than conditions under which retrieval is successful but easier. One purpose of Experiment 2 was to replicate these results. Experiment 2 was also designed to accomplish two other goals.

The first goal relates to the ISI and criterion assumptions on which the predictions of the retrieval effort hypothesis rest. To revisit, the ISI assumption is that retrieval difficulty is greater for items correctly retrieved after more versus fewer intervening items between trials; the criterion assumption is that as the number of correct retrievals increases, the difficulty of each next correct retrieval will decrease. Although the results of previous research

Table 2
Linear, power, and exponential function fits, Experiments 1 and 2.

Equation utilized	Linear $y = cx + b$ R^2	Power $y = bx^{-c}$ R^2	Exponential $y = (b * e)^{-cx}$ R^2
Experiment 1			
Group function fits:			
Short ISI short RI	.562	.821	.572
Long ISI short RI	.820	.944	.794
Long ISI long RI	.824	.951	.791
Individual function fits:	.317	.390	.332
Experiment 2			
Group function fits:			
Short contracting ISI	.774	.923	.765
Short fixed ISI	.440	.466	.428
Long contracting ISI	.753	.951	.729
Long fixed ISI	.355	.669	.370
Individual function fits:	.302	.375	.327

reporting latency measures are consistent with these assumptions, relatively few studies have reported latency measures. Thus, to provide further evidence confirming that retrieval difficulty does differ as a function of ISI and criterion level, in Experiment 2, we recorded *first key press latency* (the amount of time between presentation of the cue and when a participant began typing the answer).

Regarding the evidence available from prior research, Karpicke and Roediger (2007a) presented word pairs for three practice test trials and varied the number of intervening items between trials (massed ISI, 0–0–0; expanding ISI, 1–5–9; and equal ISI, 5–5–5). They reported response latency (amount of time from cue onset until the end of recall) for practice test trials in which items were correctly retrieved. Latencies were longer when the ISI between trials was longer (e.g., after five versus zero intervening items), and latencies decreased as the number of correct retrievals increased. Although these results are suggestive, their analyses of latency for correct retrievals were post hoc—all items received three practice trials but items may or may not have been correctly retrieved during those trials. Given that items were not assigned to criterion level, the results may have been due in part to item difficulty effects. In the current experiment, manipulation of criterion level permitted examination of differences in latency without concerns for item difficulty effects. Experiment 2 also extended beyond the results of Karpicke and Roediger (2007a) by evaluating latencies for longer ISIs and higher-criterion levels.

The second goal relates to a potential alternative interpretation of the curvilinear relationship observed between criterion level and final test performance in Experiment 1. Specifically, the pattern of results we have taken as support for the criterion prediction may have been due in part to a decreasing ISI later in practice. As items began reaching criterion, they were dropped from further test-restudy practice. Toward the end of practice, items in the higher-criterion levels were being tested with a contracting ISI because fewer items from the lower criterion levels remained to serve as fillers between each next test-study trial for these higher-criterion items. Thus, although the short and long ISIs were six and 34 items at the outset of practice in Experiment 1, by the end of practice the functional ISI was much shorter. Although a contribution of contracting ISI to the curvilinear pattern would still be consistent with the retrieval effort hypothesis in that it would provide further support for the ISI prediction, it would weaken the evidence supporting the criterion prediction. To evaluate the extent to which ISI may have contributed to the pattern of diminishing returns (which was interpreted as a criterion effect, due to less difficult retrieval as the number of correct retrievals during practice increased), in Experiment 2 we implemented comparison schedules for the short and long ISI groups in which the functional ISI was held constant throughout practice for higher-criterion items. The functional ISI for these *fixed ISI* groups in Experiment 2 was four and 29, respectively.

Method

Participants and design

Ninety-eight undergraduates from Kent State University participated in return for either course credit or pay. Criterion level (1, 3, 5, 6, 7, 8, 10) was a within-participant variable. There were a total of four groups (short contracting ISI, short fixed ISI, long contracting ISI, and long fixed ISI) with 23–25 participants in each. Data from one participant were excluded from analyses because the participant learned fewer than half of the items to criterion during practice.

Materials

Materials included 70 Swahili-English translation word pairs, as in Experiment 1. Ten items were assigned to each of seven lists, with an equivalent range of item difficulty across the lists. Assignment of list to criterion level (1, 3, 5, 6, 7, 8, or 10) was counterbalanced across participants; thus, each criterion level had ten items, as in Experiment 1.

Procedure

All task instructions and items were presented via computer. All items received initial study followed by test-restudy practice trials as in Experiment 1. The computer recorded the number of times each item was correctly retrieved, as well as the *first key press latency* for each test trial. The first key press latency was the amount of time between the cue onset and the first key pressed by the participant when entering a response in the recall field. Items continued to be practiced until they reached their assigned criterion level of performance, at which point they were dropped from further test-restudy practice. As in Experiment 1, participants were aware at the outset that they would be tested on items until they reached an "acceptable level of performance" but were not specifically aware of the number of times each item needed to be correctly retrieved.

The primary difference between the contracting and fixed ISI groups was the order of presentation of items. The presentation schedules for the contracting ISI groups (short contracting ISI and long contracting ISI) were identical to those for the short ISI and long ISI groups in Experiment 1. For the fixed ISI groups (short fixed ISI and long fixed ISI), the order of presentation was arranged so that the higher-criterion items were learned first. We adopted this presentation schedule to avoid the use of filler items

to maintain fixed ISIs for higher-criterion items. The important objective of Experiment 2 was to provide the higher-criterion levels in the fixed groups as much of an advantage as possible to determine whether diminishing returns observed in Experiment 1 were due to a contracting ISI (versus a criterion effect due to retrieval becoming less effortful as the number of correct retrievals increased). See Appendix A for a detailed description of the order of presentation for the contracting ISI and fixed ISI groups.

Following the practice phase, all groups completed a 25-min reading comprehension filler task (as in Experiment 1) that was not related to the main experimental task. Upon completion of the filler task, participants completed a self-paced cued-recall final test for all items, as in Experiment 1.

Results and discussion

The primary goal of Experiment 2 was to replicate and extend the results of Experiment 1. Below, we first report the first key press latency results that confirm the two assumptions on which the predictions of the retrieval effort hypothesis rest. We then turn to final test performance, to test the ISI and criterion predictions and to evaluate whether the same pattern of results observed in Experiment 1 replicated in both the contracting and fixed ISI groups. 95% confidence intervals (CI) around the difference between means (M_d) are reported for paired comparisons of interest below.

First key press latencies

We report mean first key press latency (in seconds) as a function of the *n*th trial on which an item was correctly retrieved for each group in Fig. 2. Results are collapsed across criterion level, because repeated measure ANOVAs indicated no significant differences in latency as a function of criterion level (e.g., as would be expected, latency for the first correct retrieval of an item did not differ as a function of the criterion level to which that item had been assigned).

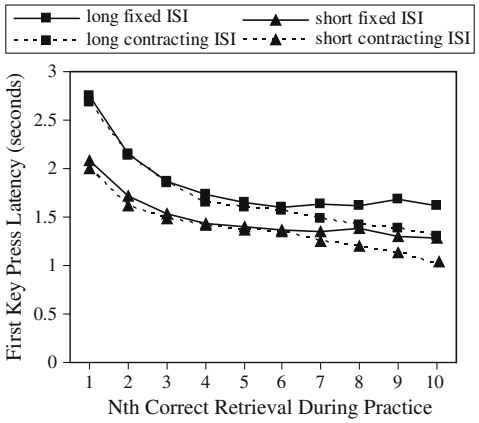

Fig. 2. Mean first key press latency in seconds for the *n*th correct retrieval during practice for each group in Experiment 2.

To test the ISI and criterion assumptions, we conducted a 2 (contracting ISI versus fixed ISI) × 2 (short ISI versus long ISI) × 10 (*n*th correct retrieval) mixed factor ANOVA. We report all results in Table 3 for completeness but will only discuss effects of interest for testing the assumptions here. First, consistent with the ISI assumption, the main effect of ISI was significant. First key press latencies were longer for correct retrievals after a longer versus a shorter ISI (long ISI, $M = 1.75$, $SE = .04$; short ISI $M = 1.44$, $SE = .03$), indicating that more effort was expended correctly retrieving items with a longer versus a shorter ISI.

Second, consistent with the criterion assumption, the main effect of the *n*th correct retrieval was significant. As the number of correct retrievals increased, first key press latencies decreased. To further evaluate the criterion assumption, for each group, we compared latencies for the first versus tenth correct retrieval. First key press latencies were shorter on the tenth correct retrieval compared to the first correct retrieval for all groups (short contracting ISI: $M_d = .98$, CI = .83, 1.14; short fixed ISI: $M_d = .80$, CI = .66, .93; long contracting ISI: $M_d = 1.39$, CI = 1.24, 1.53; long fixed ISI: $M_d = 1.18$, CI = 1.03, 1.32). Thus, retrieval became less difficult as the number of correct retrievals increased.

Although not of primary interest for evaluating the ISI and criterion assumptions, we comment briefly on the significant interaction of *n*th correct retrieval and contracting/fixed ISI. Latency differences between the contracting and fixed ISI groups emerged around the seventh correct retrieval ($M_d = .19$, CI = −.002, .24; 8th correct retrieval: $M_d = .18$, CI = .06, .31; 9th correct retrieval: $M_d = .22$, CI = .07, .38; 10th correct retrieval: $M_d = .29$, CI = .15, .43). Shorter first key press latencies for the contracting ISI groups compared to the fixed ISI groups is not overly surprising given that the ISI between correct retrievals for higher-criterion levels was shorter in the contracting group than in the fixed group toward the end of practice.

Taken together, the above analyses support both the ISI and criterion assumptions on which the predictions of the retrieval effort hypothesis rest. Next, we discuss the results for each prediction in turn.

Final test performance

The mean proportion of items correctly recalled for each criterion level at final test across participants in each group is presented in Fig. 3. The results of a 2 (contracting

Table 3
2 × 2 × 10 Mixed ANOVA results for first key press latency, Experiment 2.

	df	*F*	*p*
Main effects			
*n*th Correct retrieval	4.6, 426.9[a]	338.6	<.001
Contracting/fixed ISI	1, 92	3.31	.07
Short/long ISI	1, 92	35.5	<.001
Interactions			
*n*th Retrieval * contracting/fixed ISI	4.6, 426.9[a]	7.26	<.001
*n*th Retrieval * short/long ISI	4.6, 426.9[a]	18.04	<.001
*n*th Retrieval * contracting/fixed ISI * long/short ISI	4.6, 426.9[a]	.61	.68

[a] *Note:* Greenhouse–Geisser corrections are reported to correct for the violation of the assumption of sphericity.

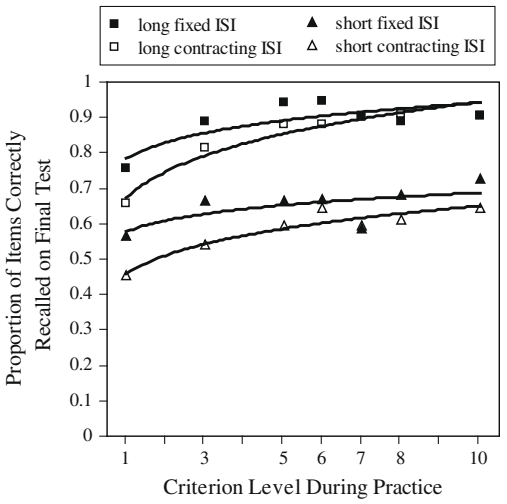

Fig. 3. Mean proportion of items correctly recalled on the final cued-recall test for each group in Experiment 2. Data points represent mean proportion of items correctly recalled at each criterion level. Lines represent the best power function fit for each group across criterion levels.

ISI versus fixed ISI) × 2 (short ISI versus long ISI) × 7 (criterion level) mixed factor analysis of variance (ANOVA) are reported in Table 4. As in Experiment 1, the general pattern of results supports the ISI prediction. A significant main effect of ISI indicated that final test performance was greater after longer versus shorter ISIs between correct retrievals (long ISI M = .87, SE = .03; short ISI M = .62, SE = .02). Overall, longer ISIs led to more difficult retrieval (as evidenced by longer latencies), which in turn led to higher levels of final test performance.

We next evaluated the criterion prediction. As in Experiment 1, we observed a pattern of diminishing returns (i.e., decreasing incremental benefit for final test performance as the number of correct retrievals increased). We evaluated whether a linear or curvilinear function fit the data best both at the group and individual level (see Table 2 for function fits). In each of the four groups, the best fitting curvilinear function (the power function) fit the data better than the linear function, accounting for 75% versus 58% of the variance across groups, respectively. To determine whether the pattern observed at the group-level analyses also held at the individual level, we computed the mean across individual R^2 values for each function fit (see Table 2). Once again, the power function provided a better fit than the linear function (M_d = .07, CI = .04, .10). Taken together, these analyses indicate that the relationship between criterion level and final test performance is curvilinear, which supports the criterion prediction of the retrieval effort hypothesis.

As described earlier, our assumption is that criterion level influenced retrieval difficulty, which in turn led to differences in final test performance. However, does criterion level per se have an effect on final test performance, above and beyond its influence on retrieval difficulty? To evaluate this relationship further, we conducted hierarchical regression analyses (see Table 5). When criterion level was entered alone as a predictor of final test performance (step 1 in Table 5), it accounted for a significant amount of variance in final test performance. However, when criterion level was entered after first key press latency (step 2 in Table 5), criterion level no longer accounted for a significant amount of variance in final test performance. These results provide converging evidence that the difficulty of correctly retrieving items (as measured by first key press latency) influences the memorial benefits of correct retrievals.

The final set of analyses concerned the extent to which the diminishing returns observed in final test performance in Experiments 1 and 2 were due to less effort expended correctly retrieving items because of a contracting ISI rather than criterion level. Inconsistent with this alternative, however, the main effect of fixed versus contracting ISI was not significant (Table 4). Furthermore, although the interaction between ISI and criterion level approached significance, this was primarily due to differences between groups for the lower criterion levels. The function fit analyses indicated that a curvilinear function fit the data best in both the fixed groups as well as the contracting groups. The potential ceiling effects in the long ISI groups warrant some caution in interpretation of the similar patterns for the contracting and fixed ISI schedules. However, the same pattern of results is observed in the short ISI groups, in which there are no ceiling effects, providing stronger evidence against the ISI interpretation of the pattern of diminishing returns. Furthermore, although the latency results indicated differences between contracting ISI and

Table 4
2 × 2 × 7 Mixed ANOVA results for final test performance, Experiment 2.

	df	*F*	*p*
Main effects			
Criterion level	4.7, 439.5[a]	22.80	<.001
Contracting/fixed ISI	1, 94	2.27	.14
Short/long ISI	1, 94	51.32	<.001
Interactions			
Criterion ∗ contracting/fixed ISI	4.7, 439.4[a]	2.16	.06
Criterion ∗ short/long ISI	4.7, 439.4[a]	2.09	.07
Criterion ∗ contracting/fixed ISI ∗ long/short ISI	4.7, 439.4[a]	.88	.49

[a] *Note:* Greenhouse–Geisser corrections are reported to correct for the violation of the assumption of sphericity.

Table 5
Summary of hierarchical regression analyses for variables predicting final test performance.

Statistics	R^2	β	R^2	β	R^2	β	R^2	β
	Short contracting		Long contracting		Short fixed		Long fixed	
Step 1								
Criterion level	.04	.19**	.17	.41***	.02	.15	.05	.22**
Step 2								
First key press latency	.04	−.12	.27	−.44***	.18	−.49***	.09	−.31***
Criterion level	.05	.12	.28	.12	.19	−.11	.09	.01

*** p < .001, ** p < .01.

fixed ISI groups at higher-criterion levels, it is important to note that final test performance reached asymptote earlier than the point at which the two groups began to differ in retrieval difficulty.

General discussion

Overall, the pattern of results from Experiments 1 and 2 confirmed predictions from the retrieval effort hypothesis, which states that successful but difficult retrievals will be better for memory than successful but easy retrievals. Specifically, both ISI and criterion level manipulations influenced the difficulty of successful retrieval during practice, which in turn led to differences in final test performance. Experiment 2 extended beyond the results of Experiment 1 by providing a measure of retrieval difficulty (first key press latency). Shorter latencies were observed for the shorter versus longer ISI group, which showed lower levels of final test performance. Additionally, latencies decreased with each next correct retrieval, as did the gain in final test performance.

Testing the predictions of the retrieval effort hypothesis not only makes an important theoretical contribution to the retrieval practice literature (in which few direct tests of a priori predictions have been conducted), it also revealed empirical patterns of results that will be useful for constraining theory. No previous research has manipulated the ISI between correct retrievals during practice. Additionally, no previous research has documented incremental gain in final test performance using different criterion levels during practice. Although the results of Experiments 1 and 2 bear surface similarity to results from previous studies, they differ in important ways. First, whereas the majority of previous studies have manipulated the ISI between *practice trials* with items (e.g., Karpicke & Roediger, 2007a; Pyc & Rawson, 2007), we manipulated the ISI between *correct retrievals* with items. Manipulating ISI between a fixed number of trials leads to a mixture of correctly and incorrectly retrieved items across practice trials, whereas manipulating the ISI between correct retrievals (criterion levels) ensures that all items are correctly retrieved the same number of times during practice. Manipulating the ISI between correct retrievals allows us to evaluate differences in final test performance as a function of criterion level without the concern of item difficulty effects.

Second, regarding the similarity of the current work to the overlearning literature, in the current experiments, we manipulated criterion level during practice. In contrast, most overlearning studies have manipulated the amount of time spent studying or number of practice trials items receive (for a review see Driskell, Willis, & Copper, 1992). *Duration-based* procedures (Rohrer et al., 2005) predetermine the number of practice trials each group receives. In *criterion-based* procedures, items are practiced until they reach a specific criterion after which the amount of additional practice is manipulated. Criterion-based procedures can be broken down further into those in which items receive either fewer or more additional practice trials (e.g., learn items to a criterion of two correct retrievals and then receive either ten or 20 additional practice trials, as in Kratochwill et al., 1977), or procedures in which items continue to be practiced until they reach a specified criterion (1, 2, or 4 correct retrievals, as in Nelson et al., 1982). To our knowledge, only one overlearning study (e.g., Nelson et al., 1982) prior to the current work has used this particular criterion-based procedure, with the majority of overlearning studies using a duration-based procedure.

The current research extends beyond previous overlearning studies in two key ways. First, we directly manipulated the ISI between correct retrievals (which to our knowledge has not previously been explored in relation to overlearning). Second, we included a wider range of criterion levels. Nelson et al. (1982) found greater memorial benefits for correctly retrieving items more compared to fewer times (as we did for lower criterion levels), but because we utilized a wider range of criterion levels (up to 10), we observed a pattern of diminishing returns. Similarly, although some duration-based overlearning studies functionally had multiple correct retrievals in overlearning conditions, these studies have not documented incremental gain in performance at different criterion levels.

An additional empirical contribution is the first key press latency measure recorded for each correct retrieval (in Experiment 2). Very few studies on retrieval practice have reported latency measures (Karpicke & Roediger, 2007a reported a total recall time measure for correct retrievals). The present study illustrates how inclusion of latency measures can inform hypothesis testing and can provide a more complete understanding of the effects of retrieval practice.

Most researchers would agree that retrieval practice benefits memory. However, not all retrieval practice produces the same benefits. Our results suggest that difficult correct retrievals are more desirable than easier ones for promoting memory. Why do more difficult retrievals benefit memory? One possible account comes from recent work by Pavlik and Anderson (2008). Based on the declarative memory component of the ACT-R architecture (Anderson & Lebiere, 1998), Pavlik and Anderson developed a model to predict the optimal spacing of repeated retrieval practice for promoting memory. Of greatest relevance here, their model assumes that each time an item practiced, the activation of that item in declarative memory receives an increase in strength that then decays as a power function of time. Each strengthening event has its own decay rate, and the activation of an item in memory at time t is the sum of the remaining strength from all prior strengthening events. Importantly, the model assumes that "higher activation at the time of a practice will result in the benefit of *that* practice decaying more quickly...if activation is low, decay will proceed more slowly" (p. 115, Pavlik & Anderson, 2008). If we add the relatively straightforward assumption that a more difficult retrieval reflects a lower activation level at practice, this model would suggest that more difficult retrievals lead to better memory by retarding forgetting.

In addition to the strength-based model discussed above, other process-based accounts may help explain why more difficult correct retrievals promote higher levels of memory. For example, McDaniel and Masson

(1985) tested various explanations for how a practice test influences the original memory representation of an item. Of interest here, unrelated concrete nouns were initially studied with either a semantic encoding task (e.g., does hawk belong to the same category as blackbird?) or a phonemic encoding task (e.g., does hawk rhyme with talk?). Participants in the retrieval practice group then received a practice cued-recall test, whereas participants in the control group did not. All participants returned 24 hours later for a final cued-recall test, in which target items were cued with semantic or phonemic cues. Final test performance was greater in the retrieval practice group than in the control group when the encoding task and the final recall cue were different (e.g., when an item was encoded in the semantic task but then cued phonemically at final test). McDaniel and Masson concluded that the practice test "served to increase the encoding variability of a target word, thereby improving its chances of delayed retrieval when the delayed cue represented different attributes from those emphasized in the original encoding" (p. 377). With respect to the current results, one possibility is that difficult retrievals enhance encoding variability to a greater extent than easier retrievals; difficult retrieval may involve the activation of more related information due to a more elaborate memory search to retrieve the target.

Further exploration of the ideas discussed above will be an important direction for future research on the relationship between retrieval difficulty and subsequent memory. Importantly, the present research provides a basic methodological approach for further exploring the retrieval effort hypothesis, as well as foundational results to support it. In supporting the predictions of the retrieval effort hypothesis, the current studies also add to accumulating support for the desirable difficulty framework. As is the case with any architecture or framework, the strength of evidence for the framework comes from an accumulation of evidence for specific hypotheses or models that follow from the framework within particular task domains. The current work adds to other task domains providing support for the desirable difficulty framework (e.g., spacing of practice, contextual interference; see Bjork, 1994 for additional examples). More generally, given that the extant literature on retrieval practice effects is largely empirical, the current work contributes to the incipient movement of this area toward more theoretical investigations into the nature of retrieval practice effects.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, US Department of Education, through Grant #R305H050038 to Kent State University. The opinions expressed are those of the authors and do not represent views of the Institute or the US Department of Education.

Thanks to Tina Burke, Sean Burton, Jill Peterson, and Ericka Schmitt for assistance with data collection. Thanks also to Heather Bailey and Nic Wilkins for assistance with data analyses. A special thanks to John Dunlosky and RADlab for helpful comments on earlier versions of this paper.

Appendix A. Overview of presentation schedule for contracting and fixed ISI groups in Experiment 2

A.1. Short contracting ISI

For illustrative purposes, suppose that items were assigned to criterion level as such (across participants, assignment of items to criterion level was counterbalanced, as described in methods above):

	Item number
Criterion 1	1, 8, 15, 22, 29, 36, 43, 50, 57, 64
Criterion 3	2, 9, 16, 23, 30, 37, 44, 51, 58, 65
Criterion 5	3, 10, 17, 24, 31, 38, 45, 52, 59, 66
Criterion 6	4, 11, 18, 25, 32, 39, 46, 53, 60, 67
Criterion 7	5, 12, 19, 26, 33, 40, 47, 54, 61, 68
Criterion 8	6, 13, 20, 27, 34, 41, 48, 55, 62, 69
Criterion 10	7, 14, 21, 28, 35, 42, 49, 56, 63, 70

For this hypothetical participant, order of item presentation would be:

Study Items 1–7 (in random order)
Retrieval practice with items 1–7 until all items reach criterion
Study Items 8–14 (in random order)
Retrieval practice with items 8–14 until all items reach criterion
Study Items 15–21 (in random order)
Retrieval practice with items 15–21 until all items reach criterion
Study Items 22–28 (in random order)
Retrieval practice with items 22–28 until all items reach criterion
Study Items 29–35 (in random order)
Retrieval practice with items 29–35 until all items reach criterion
Study Items 36–42 (in random order)
Retrieval practice with items 36–42 until all items reach criterion
Study Items 43–49 (in random order)
Retrieval practice with items 43–49 until all items reach criterion
Study Items 50–56 (in random order)
Retrieval practice with items 50–56 until all items reach criterion
Study Items 57–63 (in random order)
Retrieval practice with items 57–63 until all items reach criterion
Study Items 64–70 (in random order)
Retrieval practice with items 64–70 until all items reach criterion

A.2. Long contracting ISI

For illustrative purposes, suppose that items were assigned to criterion level as such (across participants,

assignment of items to criterion level was counterbalanced, as described in methods above):

	Item number
Criterion 1	1–5, 36–40
Criterion 3	6–10, 41–45
Criterion 5	11–15, 46–50
Criterion 6	16–20, 51–55
Criterion 7	21–25, 56–60
Criterion 8	26–30, 61–65
Criterion 10	31–35, 66–70

For this hypothetical participant, order of item presentation would be:

Study Items 1–35 (in random order)
Retrieval practice with items 1–35 until all items reach criterion
Study Items 36–70 (in random order)
Retrieval practice with items 36–70 until all items reach criterion

A.3. Short fixed ISI

For illustrative purposes, suppose that items were assigned to criterion level as such (across participants, assignment of items to criterion level was counterbalanced, as described in methods above):

	Item number
Criterion 1	1–5, 36–40
Criterion 3	6–10, 41–45
Criterion 5	11–15, 46–50
Criterion 6	16–20, 51–55
Criterion 7	21–25, 56–60
Criterion 8	26–30, 61–65
Criterion 10	31–35, 66–70

For this hypothetical participant, order of item presentation would be:

Study item 31–35
Retrieval practice with item 31
Retrieval practice with item 32
Retrieval practice with item 33
Retrieval practice with item 34
Retrieval practice with item 35

Continue retrieval practice in this order until an item reaches criterion, at which point, the slots for this item are used instead to introduce another item (from the highest criterion level with items still to be learned). For example, for this hypothetical participant, suppose that items 31 and 35 reach criterion on their 11th practice trial and items 32 and 33 reach criterion on their 12th practice trial. The schedule would then be:

Retrieval practice with item 31 (11th Trial, reached criterion)
Retrieval practice with item 32
Retrieval practice with item 33
Retrieval practice with item 34
Retrieval practice with item 35 (11th Trial, reached criterion)
Study item 66 (New item taking over slot previously used by item 31)
Retrieval practice with item 32 (12th Trial, reached criterion)
Retrieval practice with item 33 (12th Trial, reached criterion)
Retrieval practice with item 34
Study item 67 (New item taking over slot previously used by item 35)
Retrieval practice with item 66
Study item 68 (New item taking over slot previously used by item 32)
Study item 69 (New item taking over slot previously used by item 33)
Retrieval practice with item 34
Retrieval practice with item 67
Retrieval practice with item 66
Retrieval practice with item 68
Retrieval practice with item 69
Retrieval practice with item 34
Retrieval practice with item 67 (and so on)

This process continues until all items from each criterion level reaches criterion. Items are introduced in order of criterion level, with higher-criterion levels being learned first.

A.4. Long fixed ISI

The same schedule of learning occurs as in the short fixed ISI example above with the exception that 30 items are initially studied (criterion levels 10, 8, and 7) and receive retrieval practice. Items from the remaining criterion levels are filtered in one at time as items in criterion levels 10, 8, and 7 begin reaching criterion.

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum Publishers.

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

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

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.

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

Driskell, J. E., Willis, R. P., & Copper, C. (1992). Effect of overlearning on retention. *Journal of Applied Psychology, 77*, 615–622.

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

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, and 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 and Language, 57*, 151–162.

Kratochwill, T. R., Demuth, D. M., & Conzemius, W. C. (1977). The effects of overlearning on preschool children's retention of sight vocabulary words. *Reading Improvement, 14*, 223–228.

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

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

Nelson, T. O., Leonesio, R. J., Shimamura, A. P., Landwehr, R. F., & Narens, L. (1982). Overlearning and the feeling of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*, 279–288.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 3–8.

Pavlik, P. I., Jr., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied, 14*, 101–117.

Pyc, M. A., & Rawson, K, A. (submitted for publication). Costs and benefits of dropout schedules of test-restudy practice: Implications for student learning.

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition, 35*, 1917–1927.

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

Rohrer, D., Taylor, K., Pashler, H., Wixted, J., & Cepeda, N. (2005). The effect of overlearning on long-term retention. *Applied Cognitive Psychology, 19*, 361–374.