Metacognitive control in self-regulated learning: Conditions affecting the choice of restudying versus retrieval practice

Thomas C. Toppino¹ · Melissa H. LaVan¹ · Ryan T. Iaconelli¹

Published online: 29 May 2018
© Psychonomic Society, Inc. 2018

Abstract

In two experiments, learners studied word pairs one or two times and took a final cued recall test. They studied each pair upon its initial presentation and decided whether they would restudy it later, take a practice test on it later (retrieval practice), or forego all further practice with the pair. Whether learners preferred restudying or testing depended upon conditions. Regardless of whether practice tests were followed by feedback, they chose to take practice tests relatively more often when items were easy and the lag or spacing interval between the first and second occurrence was short, whereas they chose to restudy relatively more when items were hard and the lag was long. That is, they preferred testing under conditions in which successful retrieval on the practice test was likely. In Experiment 2, we varied the number of points each item was worth if recalled on the final test. A high point value led to a marked increase in both the preference for testing when the lag was short and the preference for restudying when the lag was long. Results support the hypothesis that learners appreciate at some level that retrieval practice can be a more effective learning strategy than restudying. However, they appear to believe that successful retrieval is necessary to reap the benefits of retrieval practice. As a consequence, their tendency to choose testing is influenced by conditions (item difficulty and spacing interval) that affect the likelihood of successful practice-test retrieval.

Keywords metamemory

The testing effect refers to the fact that engaging in retrieval practice on previously studied information (e.g., by taking a practice test) serves as a potent learning event that facilitates subsequent retention more than simply restudying the information (Karpicke & Roediger, 2008; Roediger & Butler, 2011; Rowland, 2014). However, learners frequently fail to recognize the beneficial effect of testing, erroneously ascribing greater benefits to restudying the information (e.g., Karpicke & Roediger, 2008). They also tend to underuse retrieval practice when they regulate their own learning (e.g., Karpicke, 2009; Karpicke, Butler, & Roediger, 2009). The present research explores the roles that spacing (or lag) and item difficulty may play in influencing the extent to which learners use practice tests during self-regulated learning.

Although tests often are conceptualized merely as tools for evaluating how well information has been learned and

retained, there were early indications that tests are also learning events (e.g., Abbott, 1909; Gates, 1917), and, over the years, some investigators have maintained a focus on the role of retrieval in modifying memory (e.g., Bjork, 1975). However, recent widespread interest in the role of retrieval practice in learning and retention was spawned by two seminal papers by Roediger and Karpicke (2006a, 2006b). The first was a powerful empirical demonstration of the testing effect, using what has become a more or less standard experimental design. Roediger and Karpicke's (2006a) participants studied prose passages and then either restudied the information one or more times or took an equivalent number of practice tests (free recall) without receiving feedback. Thus, any advantage of testing could not be attributed merely to additional exposure. A final free-recall test showed that, although restudying produced a small performance advantage after a 5-min retention interval, taking practice tests produced a large performance advantage after retention intervals of 2 days and 1 week.

Roediger and Karpicke (2006b) provided an informative review of the empirical literature and a critical analysis of theoretical alternatives. They distinguished between direct (unmediated) effects of testing and indirect (mediated) effects.

✉ Thomas C. Toppino
thomas.toppino@villanova.edu

¹ Department of Psychological and Brain Sciences, Villanova University, 800 Lancaster Avenue, Villanova, PA 19085, USA

Direct effects refer to the beneficial effects on learning that result from the act of retrieval per se. The clearest evidence for direct effects of testing comes from the fact that practice tests without feedback or subsequent study opportunities result in better learning and retention than restudying the information. Indirect effects refer to situations in which testing “promotes learning via some other process or processes” (p. 182). For example, practice tests can help learners distinguish what they know from what they do not know so they can more effectively focus further studying. Also, when practice tests are followed by feedback that provides the correct answer or by additional study opportunities, prior retrieval may have the indirect effect of enhancing or potentiating subsequent encoding (e.g., Arnold & McDermott, 2013; Izawa, 1966). Testing may potentiate subsequent learning even if the tested information was not studied previously (e.g., Hays, Kornell, & Bjork, 2013).

A wealth of evidence now exists supporting the robustness of the testing effect (see Rowland, 2014, for a meta-analytic review). The phenomenon has been obtained both in the laboratory (e.g., Roediger & Butler, 2011; Roediger & Karpicke, 2006a; Rowland, 2014; Toppino & Cohen, 2009) and in educational settings (e.g., McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; McDermott, Agarwal, D’Antonio, Roediger, & McDaniel, 2014). The benefit of testing tends to be greater when practice tests involve free recall or cued recall rather than recognition (but see Little & Bjork, 2012; Little, Bjork, Bjork, & Angello, 2012) and does not depend on the same format being used for both the practice and final tests (e.g., Carpenter & DeLosh, 2006; Glover, 1989). The effect is often greater after a long retention interval, as was found by Roediger and Karpicke (2006a; see also Toppino & Cohen, 2009; Wheeler, Ewers, & Buonanno, 2003), but the benefit of retrieval practice frequently has been obtained with short retention intervals as well (e.g., Carpenter, 2009; Kuo & Hirshman, 1996; Rowland & DeLosh, 2015).

Despite the large benefit to learning and retention that practice tests provide relative to restudying the target information, learners’ metacognitive awareness of this benefit seems quite limited, in terms of both monitoring the effectiveness of testing and controlling learning strategies during self-regulated learning. Regarding monitoring, when learners are asked to predict their future retention (make a judgment of learning or JOL) after restudying or taking a practice test, they frequently predict that they will remember restudied material better than tested material, although they actually remember tested material better (e.g., Kornell & Son, 2009, Experiment 1; Roediger & Karpicke, 2006a). Thus, when people monitor the extent of their learning after restudying or taking a practice test, they exhibit little awareness that they learned more from retrieval practice, although providing learners with experience and appropriate external support for their metacognitive assessments

can significantly enhance their sensitivity to the effects of testing (Tullis, Finley, & Benjamin, 2013).

A comparable issue involving metacognitive control concerns the extent to which individuals fail to appreciate the potential benefits of retrieval practice when they choose study strategies during self-regulated learning. Monitoring and control findings in related domains have not always matched. For example, in the area of distributed practice, monitoring research has indicated little or no awareness that spaced practice provides greater learning benefits than massed practice does (e.g., Baddeley & Longman, 1978; Logan, Castel, Haber, & Viehman, 2012; Zechmeister & Shaughnessy, 1980), whereas learners consistently choose spaced over massed practice in self-regulated learning as long as the study durations are sufficient for them to perceive the stimuli (e.g., Ariel, Dunlosky, & Toppino, 2014; Benjamin & Bird, 2006; Pyc & Dunlosky, 2010; Toppino & Cohen, 2010; Toppino, Cohen, Davis, & Moors, 2009).

Several sets of researchers have investigated metacognitive control by administering surveys asking people about the study strategies they use in real-world learning situations (e.g., Hartwig & Dunlosky, 2012; Karpicke et al., 2009; Kornell & Bjork, 2007; Yan, Thai, & Bjork, 2014). In all cases, restudying material was a strategy reported by a high proportion of respondents. The proportion who reported self-testing varied considerably but was surprisingly high in some cases. Both Kornell and Bjork (2007) and Hartwig and Dunlosky (2012) found that approximately 90% of their survey respondents reported using some form of self-testing when they study. However, the majority of respondents reported doing so in order to diagnose their level of learning.¹ Only a minority (18% in the Kornell & Bjork study and 27% in the Hartwig & Dunlosky study) indicated that they believed that testing led to better learning. These findings suggest that although learners may recognize at least some of the indirect benefits of retrieval practice, they rarely appreciate its direct benefits. However, confidence in this conclusion should be tempered by the fact that surveys are relatively crude instruments for assessing learning processes and may be especially insensitive to the potential effect of situational influences on the learners’ choice of strategies.

Is it possible that learners appreciate the direct benefit of retrieval practice more than the survey data would suggest? The clearest evidence that learners might have some awareness of the direct benefits of testing would involve showing that they prefer taking practice tests over restudying under conditions in which no feedback is provided on the practice tests. This set of conditions, however, may occur only rarely during the real-world studying addressed by surveys. The extent to which learners appreciate the direct benefits of testing

¹ A similar conclusion was reached by Wissman, Rawson, and Pyc (2012) in a survey of how and why students study with flashcards.

may be more readily addressable in laboratory studies in which learners' choice of restudying or taking practice tests can be observed while they are actively engaged in learning and in which situational factors can be manipulated. Surprisingly, we could find few studies that took this approach. The studies that did (e.g., Karpicke, 2009; Kornell & Son, 2009) significantly advanced our knowledge of metacognitive control over retrieval practice but, from our perspective, they left lingering questions, especially regarding learners' appreciation of the direct benefits of practice tests. We were concerned that when no feedback is provided following practice tests, learners may reasonably assume that successful retrieval is necessary for a practice test to have a beneficial effect, an assumption shared by researchers and theorists (e.g., Kang, McDermott, & Roediger, 2007; Roediger & Butler, 2011; Storm, Friedman, Murayama, & Bjork, 2014). Thus, they may be more likely to prefer a practice test over restudying when the to-be-learned material is perceived to be easy or relatively well learned. They also may prefer testing when the interval or lag between studying and the practice test is relatively short, increasing the likelihood of successful retrieval. Fear of retrieval failure with harder items and longer lags may deter learners from taking practice tests.

Karpicke (2009) allowed participants some control over their strategies in two of the four experiments he reported, but only Experiment 2 addressed the direct effect of testing. In that experiment, participants received alternating study and test trials on 60 Swahili–English word pairs until all items had been recalled at least once. When a given item was recalled initially, participants chose from one of three strategies. They could study it for two more trials but not take any further tests. They could test it on two more test trials but not study it again. Or they could drop the pair from future trials (no further studying or testing). Results indicated a trend for participants to choose practice tests (without feedback) more often than restudying, but the small difference was dwarfed by a strong preference to simply drop items from further practice once they had been recalled. The fact that dropped items were recalled very poorly on the final test indicates that memory for those items would have benefited from additional practice of either kind. Participants greatly overestimated how well they knew a pair after they recalled it for the first time, apparently believing that they would recall it on the final test even without further practice. Although participants chose to be reexposed to an item less than half the time, the weak tendency for testing without feedback to be preferred over restudying is potentially consistent with some appreciation of the direct benefit of testing. This preference may have been weakened to the extent that testing was discouraged by the long lags that intervened on the average between repetitions of an item on successive trials.

Kornell and Son (2009), Experiment 2 conducted an experiment in which they asked participants to learn four successive 12-pair lists for a final cued-recall test that was administered after all lists had been studied. After initially studying a list, participants chose whether to study the entire list a second time or to take a practice test on the entire list. Participants either received feedback during the tests (i.e., the correct answer was presented for 1 s after participants entered a response) or received no feedback, depending on the group to which they were assigned. Results indicated that, regardless of the availability of feedback on tests, participants initially had no preference for testing, choosing to restudy and to take practice tests equally often on the first list. A preference for taking practice tests emerged with experience over successive lists, but the evidence was equivocal in the critical no-feedback condition. Although the tendency to select testing over restudying did not differ reliably as a function of feedback, the preference for testing was greater in the feedback condition and failed to exceed chance in the no-feedback condition on any trial. The ambiguity may be tied to issues of statistical power, but it remains unclear whether experience over successive lists in the no-feedback condition was sufficient for learners to adopt a testing strategy that might indicate a growing appreciation of the benefits of testing per se (i.e., the direct effect of testing). A concern is that the lag between an item's presentation on the first and second trial of a given list was relatively long on the average and could have deterred the choice of testing when feedback was not provided.

The present research sought to clarify the circumstances under which testing might be preferred to restudying during self-regulated learning, especially when practice tests are not followed by feedback. We expected that the extent to which learners chose practice testing would depend on the degree to which they expected to be able to retrieve successfully on the tests. Thus, we expected testing to be chosen relatively more often when items tended to be easy (i.e., more memorable) and when the lag between initial study and test tended to be short. Our interest was not in whether learners adopted an optimal strategy. For example, the relationship between lag and practice tests is complex. In some cases, longer lags may be more beneficial than shorter lags (e.g., Pyc & Rawson, 2009; Rawson, Vaughn, & Carpenter, 2015), whereas, in other cases, shorter lags may be more effective (e.g., Maddox, Balota, Coane, & Duchek, 2011; Storm, Bjork, & Storm, 2010). Our question was whether there are conditions in which learners *believe* they will learn better through retrieval practice than through restudying. Most observers have concluded that learners do not engage in retrieval practice as much as they should and that they fail to do so because they do not appreciate that retrieval practice leads to better retention than restudying. However, to the extent that learners do appreciate the benefits of engaging in retrieval practice, their failure to use it optimally may be attributable to other factors. We will return to this issue in the General Discussion.

Our experiments combined features of the prior work by Kornell and Son (2009) and Karpicke (2009). Like Kornell and Son, we presented multiple lists of word pairs in which each pair could be presented twice, and we assessed memory on a cued-recall test after all lists had been presented. The first presentation of a pair was always a study trial, after which learners made a strategy decision that affected its second occurrence. Like Karpicke, our participants made strategy decisions on an item-by-item basis, choosing to restudy an item, take a practice test on it, or drop it from further practice. If an item were selected for further practice, the restudy trial or practice test occurred either after a lag of two other item presentations or after all items in the current list had been presented at least once. We chose a lag of two for the short lag condition because it afforded a short retention interval in the case of a practice test without being trivial. The longer lag condition followed a standard procedure introduced by Son (2004) in the course of investigating the metacognitive control of spaced practice. We also varied the difficulty of the pairs in which the member words were either weakly associated (easy) or unrelated (hard) to ensure some variability among the items in terms of difficulty.

Experiment 1

On the initial presentation of a word pair, learners chose to restudy it, take a practice test on it, or forego further practice with it. The proportion of trials on which each strategy was selected constituted the primary data of this experiment. We were especially interested in how learners' choices would be affected by the spacing separating the first study opportunity and its second presentation and by the difficulty of the item as reflected by JOLs made after the initial study opportunity. We also varied whether or not practice tests (if they were selected) were followed by feedback in which the complete pair was presented. This was motivated by the fact that Kornell and Son's (2009) results seemed inconclusive with respect to the effect of feedback versus no feedback on learners' strategy choices.

Method

Design and participants The experimental design was a $2 \times 2 \times 3$ mixed factorial with the first two factors (spacing: short or long and feedback: feedback or no feedback) manipulated between participants and the last two factors (item type: associated or unrelated, and list: 1, 2, or 3) varied within participants. Based on previous research in our laboratory, we targeted a sample size of 96 introductory psychology students. However, the data from one participant had to be discarded for noncompliance with instructions. This resulted in a final

sample of 95 participants who were assigned randomly to the four between-participant conditions such that three of the groups consisted of 24 participants and one (short spacing with no feedback after practice tests) consisted of 23 participants.

Materials Lists were created from 36 word pairs. Half were relatively easy, related pairs, with forward association strengths of .050 to .054 (Nelson, McEvoy, & Schreiber, 2004). Half were more difficult, unrelated pairs with zero normative associative connection. Both easy and hard pairs were created from common words containing three to seven letters ($M_{\text{Length}} = 4.9$ letters, in both cases). Participants studied three successive 12-pair lists with the order of the lists randomly distributed across conditions. To ensure variation in item difficulty, half of the items in each list were associated and half were unrelated. However, for purposes of analyses involving item difficulty, we relied on perceived item difficulty as reflected in JOLs reported by the participants.

Procedure Participants were instructed that during the learning phase of the experiment, they would study three separate lists of 12 pairs each, with a 30-s rest period between lists. Participants received detailed instructions on JOLs and how to make them as well as on the different strategy choices they could make (i.e., restudy, test, or done). They were told to make whatever choices they believed would be most effective in fostering learning of the pairs. The goal was to be able to perform as well as possible on a cued-recall test that would be administered at the end of the learning phase of the experiment, after all three lists had been presented. Participants also were informed about their condition so that they knew whether the spacing between the first and second presentation of an item would be short or long and whether practice tests would be followed by feedback. When the spacing was short, two item presentations intervened between the first and second occurrence of an item. When the spacing was long, items chosen for restudying or testing reoccurred at the end of the current list in a new random order. Whereas this procedure led to a range of actual lags, the overall mean lag was expected to be approximately 12 intervening item presentations.

During the learning phase of the experiment, pairs were presented successively on a computer screen in an order that was randomized independently for each participant. The initial presentation of a pair consisted of the intact word pair (cue and target together). After 6 s, the word pair disappeared, and participants were prompted to make a JOL, indicating the likelihood on a scale of zero to 10 of recalling the pair in about 15 minutes. Then, participants chose whether they wanted to restudy the pair, test the pair, or be done with the pair, in which case the pair would not be seen again until the final test. Depending on their lag condition, participants received the second presentation of an item after two other pairs had been

presented (short) or after all pairs had been presented at least once (long). If they had chosen to restudy an item, the intact pair (cue and target) was presented for 8 s on the item's second occurrence. If they had chosen to take a practice test, only the cue word was presented, and participants attempted to type the corresponding target word. Participants in the feedback condition saw the cue word by itself for 6 seconds, followed by the cue and target for 2 seconds. Participants in the no-feedback condition saw only the cue word for the full 8 seconds. Following the procedure of Kornell and Son (2009), each list was not followed by a recall test. Rather, after all lists had been studied, participants completed a 5-minute distractor task involving simple math problems and then took one cued-recall test on the pairs from all three lists, randomly ordered.

Results

All reported analyses employed an alpha level of .05 unless otherwise specified.

JOLs A $2 \times 2 \times 2 \times 3$ (Difficulty $\times$ Feedback $\times$ Spacing $\times$ List) mixed ANOVA was performed on the participants JOLs. As expected, we found a main effect of item difficulty, with much higher JOLs for easy items ($M = 7.149$) than for difficult items ($M = 4.113$), $F(1, 91) = 671.091$, $MSE = 1.951$, $\eta_p^2 = 0.881$. We also found that JOLs declined over successive lists from a high in List 1 ($M = 5.734$) to a low in List 3 ($M = 5.492$), $F(2, 91) = 3.326$, $MSE = 0.903$, $\eta_p^2 = 0.035$, and that the feedback condition yielded higher JOLs ($M = 5.857$) than the no-feedback condition did ($M = 5.400$), $F(1, 91) = 5.968$, $MSE = 4.995$, $\eta_p^2 = 0.062$. No interactions were reliable.

Strategy choice Our primary interest was on how participants' strategy choices were impacted by lag, perceived item difficulty as indexed by JOLs, feedback, and successive lists. With respect to strategy choices, we analyzed only the proportion of items that participants chose to restudy and the proportion on which they chose to be tested. Although they had the option to discontinue practice with items by choosing "done," this choice was not included in the analysis for several reasons. First, we were interested in the proportion of times participants chose to restudy and to be tested when they thought there would be a benefit to engaging in further practice. The "done" option was included in the procedure because, without it, participants would be forced to select a strategy regardless of whether they expected further practice to be beneficial, and, under those circumstances, their selections might be systematically different, as Son (2004) reported. Second, the proportion of done choices is conceptually redundant. Any change in the proportion of times participants choose to be reexposed to a pair corresponds to an opposite change in the proportion of times they chose to be done. Finally, if the done option were included in the analysis, the three levels of choice would be

completely dependent (i.e., the proportion of times any two were chosen would uniquely determine the proportion of times the third was chosen). With respect to JOLs as a measure of perceived item difficulty, we followed Son (2004) by Vincentizing each participant's JOL ratings into thirds, corresponding to the lowest, medium, and highest ratings. Participants differ with respect to the absolute level of their JOLs, but we were primarily interested in the relative JOLs within participants. Vincentizing the scores effectively normalizes the JOL data.

The proportion of items chosen for restudying and practice testing were included in a $2 \times 2 \times 2 \times 3 \times 3$ (Feedback $\times$ Spacing $\times$ Choice $\times$ List $\times$ JOL) ANOVA with repeated measures on the last three factors.² The main effect of JOL was reliable, $F(2, 182) = 48.531$, $MSE = .069$, $\eta_p^2 = 0.348$, indicating that items with higher JOLs are less likely to be selected for further practice. The main effect of choice, revealing a trend for participants to prefer practice testing over restudying, approached but failed to reach conventional significance levels, $F(1, 91) = 2.699$, $MSE = 1.028$, $\eta_p^2 = 0.029$. Neither the main effects of list nor spacing approached significance. However, there were two significant three-way interactions.

The first of these interactions is a JOL $\times$ Spacing $\times$ Choice interaction, $F(2, 91) = 4.737$, $MSE = 0.169$, $\eta_p^2 = 0.049$, which is depicted in Fig. 1. To probe this interaction, we analyzed data within the short-lag and long-lag conditions separately. A 3×2 (JOL $\times$ Choice) ANOVA on the short-lag data revealed a significant JOL $\times$ Choice interaction, $F(2, 46) = 5.807$, $MSE = 0.052$. Probing further, we compared the proportion of times participants chose to restudy or test at each JOL level. For the lowest JOL items, there was no significant difference in the degree to which participants chose to restudy or test, $t(46) = 0.048$, but participants chose to test more often than to restudy when JOLs were medium or high. This tendency was marginally significant for medium JOLs, $t(46) = -1.896$, and significant for high JOLs, $t(46) = -2.727$.

A parallel 3×2 (JOL $\times$ Choice) ANOVA on the long spacing data also revealed a significant JOL $\times$ Choice interaction, $F(2, 46) = 27.151$, $MSE = 0.080$. Probing the interaction, we found that, when participants were least confident about their ability to remember (low JOLs), they chose to restudy significantly more often than to take a practice test, $t(47) = 3.045$. As JOLs increased, however, participants became more likely to choose testing rather than restudying. This effect was nearly significant at the medium JOL level, $t(47) = -1.901$, and was significant at the high JOL level, $t(47) = -3.208$.

The second three-way interaction revealed by the omnibus ANOVA is a Feedback $\times$ List $\times$ Choice interaction, $F(2, 91) = 4.388$, $MSE = 0.082$, $\eta_p^2 = 0.046$, which is displayed in Fig. 2. To probe this interaction, we analyzed the data for

² A parallel analysis substituting normative difficulty for perceived difficulty (JOLs) revealed parallel results.

Fig. 1 Mean proportion of choices (Restudy, Test, and Done) as a function of JOL and Lag in Experiment 1. Error bars depict one standard error of the mean

participants who received feedback separately from those who did not receive feedback. A 3×2 (List $\times$ Choice) ANOVA for the no-feedback condition revealed no significant effects, indicating that choices did not change significantly with experience. However, when the same analysis was performed on data from the feedback condition, we found a significant List $\times$ Choice interaction, $F(2, 47) = 7.904$, $MSE = 0.034$. We probed this interaction by comparing participants' decisions with test or restudy separately for each list. On List 1, participants exhibited no significant preference for one strategy over the other. However, after List 1, participants who received feedback were significantly more likely to choose practice testing over restudying, $t(47) = -3.868$, and $t(47) = -3.913$, for Lists 2 and 3, respectively.

We reported earlier that participants gave higher JOLs in the feedback than in the no-feedback condition, raising the question of whether higher JOLs, rather than feedback, is responsible for the List $\times$ Feedback $\times$ Choice interaction. Two additional analyses, however, suggested that feedback is the critical variable. First, by eliminating participants with extreme mean JOL scores, we identified subgroups with similar JOL levels for the feedback ($n = 36$, $M = 5.862$, $SD = 0.565$) and the no-feedback ($n = 28$, $M = 5.893$, $SD = 0.459$) groups, respectively. Analyses revealed the same pattern of results as in the full sample, including a significant List $\times$ Feedback $\times$ Choice interaction, $F(2, 62) = 3.117$, $MSE = 0.080$, $\eta_p^2 = 0.048$. Second, parallel results were obtained by an ANCOVA on the entire data set with each participant's mean JOL as the covariate.

Fig. 2 Mean proportion of choices (Restudy, Test, and Done) as a function of Feedback and List in Experiment 1. Error bars depict one standard error of the mean

Cued Recall As we noted earlier, the primary results of this experiment involved the strategy choices that participants made during the learning phase of the experiment. The final recall test provided participants with the motivation for learning, but meaningful statistical analyses of the recall results were precluded by the facts that the data were severely confounded by participants' choices and that the data of individual participants included many empty cells. Nevertheless, keeping the limitations of the data in mind, we report the overall recall means in Table 1.

Discussion

The primary results of this experiment indicate that the degree to which learners choose to restudy information or to take practice tests varies with item difficulty and the spacing between the items first and second presentations. Learners most frequently choose to restudy when they perceive items to be difficult or poorly learned and when the spacing between successive occurrences is long. The choice of testing becomes relatively more likely when items are perceived to be easy or memorable and when the lag between studying and the practice test is short. In more general terms, learners are relatively more likely to take practice tests under conditions in which retrieval is more likely to be successful.

Independent of item difficulty and spacing, learners' strategy choices were influenced by experience with successive lists and by whether practice tests were followed by feedback. Consistent with the findings of Kornell and Son (2009), we found no significant overall preference for testing or restudying on the first list, but a preference for testing emerged with experience on successive lists when practice tests were followed by feedback. Whereas Kornell and Son's results were equivocal with respect to whether a preference for testing emerged with experience when practice tests were presented without feedback, our findings revealed no hint of a change in

preference over successive lists in the no-feedback condition. When no overall preference was obtained for testing or restudying, it was attributable to two underlying effects cancelling one another out. We consistently obtained evidence for an increased relative tendency to choose testing when retrieval success on the practice test was likely (e.g., easy items and a short spacing interval) and a reduced relative tendency to choose testing when retrieval success was less likely (e.g., hard items and a long spacing interval). That pattern was evident on List 1 for both the feedback and the no-feedback conditions and did not change with further practice although the overall frequency with which testing was chosen increased over successive lists in the feedback condition. One interpretation of Kornell and Son's findings is that they happened to select a set of learning conditions (including item difficulty and spacing) that did not favor either restudying or testing and, like us, obtained little effect of experience with successive lists when feedback was not presented after practice tests.

Experiment 2

This experiment is a partial replication and extension of Experiment 1. We again varied item difficulty, spacing, and list. Item difficulty was experimentally varied, and four successive lists were used rather than three. Practice tests were never followed by feedback. We wanted to ensure that the situationally controlled preference for testing without feedback that we found in Experiment 1 could be replicated. A preference for testing when no feedback is provided may suggest some appreciation that retrieval practice per se is beneficial. Learners who have no understanding of the direct effects of retrieval practice should always choose restudying if they wish to maximize their learning when practice tests are not followed by feedback. Taking a practice test entails the risk that retrieval could fail and a restudy opportunity could be

Table 1 Proportion of correct cued recall in Experiment 1 as a function of JOL, feedback, spacing, and choice

	Low JOL		Med JOL		High JOL	
	Short spacing	Long spacing	Short spacing	Long spacing	Short spacing	Long spacing
No feedback						
Restudy	0.42	0.46	0.68	0.57	0.83	0.94
Test	0.48	0.64	0.56	0.85	0.72	0.93
Done	0.16	0.41	0.52	0.64	0.82	0.90
Feedback						
Restudy	0.38	0.58	0.59	0.93	0.91	1.00
Test	0.70	0.71	0.89	0.80	0.99	0.99
Done	0.28	0.57	0.67	0.79	0.84	0.88

Note. Due to empty cells for many individual participants, these data are the overall proportions of correct recall (number of correct recalls divided by the number of opportunities, combined across participants)

lost. However, there may be motivations for choosing to take practice tests despite some risk. For example, in surveys, learners say they frequently use tests to diagnose their level of knowledge (e.g., Hartwig & Dunlosky, 2012; Kornell & Bjork, 2007; Kornell & Son, 2009; Yan et al., 2014). Reducing uncertainty by assessing knowledge may be a potent motivator that makes taking practice tests attractive despite the risk. In this scenario, learners may be more likely to succumb to the attraction of a test for the purposes of assessment when the risk of retrieval failure is perceived to be low. If so, even learners who have no understanding of the direct effects of testing might produce the pattern of findings we obtained in Experiment 1.

To address this issue in Experiment 2, we introduced a variable that has been used successfully to set learners' agendas and control the factors on which their metacognitive decisions are based (e.g., Ariel, Dunlosky, & Baily, 2009; Toppino & Cohen, 2010). We varied the value of recalling an item on the final cued-recall test. We told our participants that their goal should be to achieve the highest point total possible on the final test and that the point value of a pair (5 points vs. 1 point) would be indicated when it was presented for study. We expected our participants' motivation to use what they regarded as the maximally effective strategy to be greater for the high-value items. Thus, for example, if the tendency to choose practice testing for easy items and short spacing intervals reflects a risky strategy by learners who do not appreciate the direct benefits of testing, we should find less testing for high-value than for low-value items. Alternatively, to the extent that learners believe testing to be a superior learning technique under certain circumstances, we might find more testing for high-value than for low-value pairs.

Method

Design and participants The experimental design was a $2 \times 2 \times 2 \times 4$ mixed factorial. The first factor (spacing: short or long) was manipulated between participants, and the last three factors (item difficulty: easy or hard, value: 1 or 5 points, and list: 1, 2, 3, and 4) were varied within participants. Compared with Experiment 1, we sought to increase the number of participants at each level of our between-participants factor (spacing). Our final sample consisted of 69 participants, one short of our targeted sample size, with 35 and 34 participants randomly assigned to the short-spacing and long-spacing conditions, respectively.

Materials Twelve word pairs were added to the 36 used in Experiment 1 to create a pool of 48 pairs with the same characteristics. Participants studied four successive 12-pair lists consisting of six easy and six hard pairs. Half of each type of pair was arbitrarily assigned a point value of either 1 or 5.

Procedure

The procedures were the same as those of Experiment 1 with the following exceptions. During learning, pairs appeared on the screen simultaneously with their associated point values. Participants were instructed that the point value represented the number of points they would receive on the final test if they correctly recalled that item. They also were instructed that their primary objective was to score as many points as possible on the final test. On the initial presentation of each pair, participants made a strategy decision (re-study, practice test without feedback, or done) but did not make a JOL. To simplify the task for the participants, we relied on the a priori normative relationship between members of each pair to determine item difficulty.³

Results

Strategy choice The results of primary interest again concerned the proportion of trials on which each strategy was selected. We conducted a $2 \times 2 \times 2 \times 2 \times 4$ (Spacing $\times$ Point Value $\times$ Choice $\times$ Item Difficulty $\times$ List) mixed ANOVA with repeated measures on the last four factors. The main effect of item difficulty was reliable, $F(1, 67) = 22.507$, $MSE = .112$, $\eta_p^2 = .251$, indicating that harder items were more likely to be selected for further practice. The main effect of Value was significant, $F(1, 67) = 11.936$, $MSE = .060$, $\eta_p^2 = .151$, indicating that high-value items were more likely to be chosen for further practice than low-value items were. And, the main effect of choice was significant, $F(1, 67) = 4.395$, $MSE = 2.034$, $\eta_p^2 = .062$, revealing that, overall, participants preferred taking practice tests to restudying. The main effects of list and spacing, however, both failed to approach significance. The ANOVA also revealed significant two-way interactions involving item difficulty and choice, $F(1, 67) = 40.637$, $MSE = .583$, $\eta_p^2 = .378$, and spacing and choice, $F(1, 67) = 8.133$, $MSE = 2.034$, $\eta_p^2 = .108$, both of which were qualified by a marginally significant Item Difficulty $\times$ Spacing $\times$ Choice interaction, $F(1, 67) = 3.528$, $MSE = .583$, $\eta_p^2 = .050$. We probed the latter interaction by assessing the Item Difficulty $\times$ Choice interaction at each level of spacing (see Fig. 3). The interaction was reliable at both levels of spacing, $F(1, 34) = 9.943$, $MSE = .075$, $\eta_p^2 = .226$, for short spacing, and $F(1, 33) = 34.703$, $MSE = .071$, $\eta_p^2 = .513$, for long spacing. These significant Item Difficulty $\times$ Choice

³ It is widely assumed that learners assess how well they know items, regardless of whether they make overt JOLs. There also is evidence that learners' JOLs generally reflect objective item difficulty (e.g., Dunlosky & Matvey, 2001; Koriati, 1997). And, in a study of metacognitive control involving the distribution of practice, Benjamin and Bird (2006) obtained comparable results when item difficulty was experimentally manipulated and when it was measured by participants' JOLs.

Fig. 3 Mean proportion of choices as a function of item difficulty and spacing in Experiment 2. Error bars depict one standard error of the mean

interactions were probed further by assessing the effect of choice at each level of item difficulty. Participants in the short-spacing condition significantly preferred testing easy items over restudying them, $t(34) = 5.66$, whereas they exhibited no reliable practice-strategy preference for hard items, $t(34) = 1.43$. Participants in the long-spacing condition also significantly preferred testing easy items over restudying them, $t(33) = 2.10$. In contrast, however, they displayed a strong preference for restudying hard items rather than taking practice tests on them, $t(33) = -3.33$.

The overall ANOVA did not reveal a significant Point Value $\times$ Item Difficulty interaction or any higher-order interaction subsuming this two-way interaction. However, there was a significant Point Value $\times$ Choice $\times$ Spacing interaction, which was qualified by a significant Spacing $\times$ Point Value $\times$ Choice $\times$ List interaction, $F(3, 201) = 3.232$, $MSE = .074$, $\eta_p^2 = .046$. To probe this four-way interaction, we assessed the Spacing $\times$ Point Value $\times$ Choice interaction at each list level (see Fig. 4). The three-way interaction was significant for List 1, $F(1, 67) = 11.746$, $MSE = .058$, $\eta_p^2 = .149$, but not for Lists 2, 3, and 4. In probing the three-way interaction in List 1, we found that the Point Value $\times$ Choice interaction was significant both for the short-spacing condition, $F(1, 34) = 6.849$, $MSE = .042$, $\eta_p^2 = .168$, and the long-spacing condition, $F(1, 33) = 5.331$, $MSE = .074$, $\eta_p^2 = .139$. For the two-way interaction in each spacing condition, we then assessed the effect of choice at each level of point value. Participants in the short-spacing condition significantly preferred to test high-valued items, $t(34) = 3.24$, whereas participants in the long-spacing condition significantly preferred to restudy high-valued items, $t(33) = -2.26$.

For low-valued items, neither spacing condition yielded a significant preference for testing or restudying.

Although the three-way interaction was not reliable in Lists 2, 3, and 4, analyses of all three lists revealed a significant Spacing $\times$ Choice interaction, $F(1, 67) = 7.721$, $MSE = .325$, $\eta_p^2 = .103$, $F(1, 67) = 6.772$, $MSE = .302$, $\eta_p^2 = .092$, and $F(1, 67) = 5.822$, $MSE = .335$, $\eta_p^2 = .080$, for Lists 2, 3, and 4, respectively. Participants were significantly more likely to choose testing over restudying in the short-spacing condition, $t(34) = 3.49$, $t(34) = 4.05$, and $t(34) = 3.22$, for Lists 2, 3, and 4, respectively. However, in the long-spacing condition, participants were less likely to choose to take a practice test and more likely to choose to restudy, resulting in no significant preference for either strategy.

Cued recall Analyses of the data from the final cued-recall test were precluded by the same limitations that characterized the recall data of Experiment 1. Nevertheless, the overall recall means are presented in Table 2.

Discussion

The findings of this experiment replicated the essential results of Experiment 1. Easier items and shorter spacing led to an increase in the degree to which participants chose to take practice tests rather than to restudy, whereas harder items and longer spacing lessened the relative tendency to choose testing and increased the relative tendency to choose restudying. The new element in Experiment 2 was that we varied the point value that each pair was worth if it were recalled on the final cued-recall test. We expected that learners

Fig. 4 Mean proportion of choices, as a function of value and spacing, for Lists 1, 2, 3, and 4 of Experiment 2. Error bars depict one standard error of the mean

would be motivated to use the strategy that they believed to be maximally effective in producing learning when the value of items was high. In particular, if learners' choice of practice testing reflected a risky strategy rather than a belief in the superiority of testing, then they should be less likely to select testing for high-valued items.

Point value proved to be an effective moderator of learning behavior in that participants chose to engage in additional practice more often with high-valued than with low-valued items. Contrary to the risky strategy hypothesis, we found that on List 1, high-value items led learners to choose testing relatively more often when the lag was short. Learners also chose

to restudy high-value items relatively more often when the lag was long. These findings suggest that, at the outset of learning, participants judged testing to be a more effective learning strategy than restudying when spacing was short while also judging restudying to be more effective than testing when spacing was long.

On Lists 2 through 4, learners continued to choose testing more often for short than for long spacing and restudying more for long than for short spacing, but the effect of point value was no longer evident. Experience with the first list apparently altered the influence of point value on subsequent lists. The fact that point value exerted a clear effect on the first

Table 2 Proportion of correct cued recall in Experiment 2 as a function of item difficulty, point value, spacing, and choice

	Easy items		Hard items	
	Short spacing	Long spacing	Short spacing	Long spacing
Low value				
Restudy	0.64	0.74	0.34	0.45
Test	0.87	0.84	0.53	0.51
Done	0.67	0.54	0.24	0.04
High value				
Restudy	0.67	0.73	0.30	0.40
Test	0.87	0.82	0.53	0.60
Done	0.73	0.60	0.43	0.06

Note. Due to empty cells for many individual participants, these data are the overall proportions of correct recall (number of correct recalls divided by the number of opportunities, combined across participants)

list when participants' experience with the experimental task was minimal suggests that our participants' practice strategies on the first list were based on their a priori expectations about the learning conditions they would encounter. These, in turn, were affected by the instructions that stressed the importance of point value and that implied that maximizing performance required applying the more effective strategy to the high-value items. Once the participants had experienced our task, however, their strategy shifted on Lists 2 through 4 such that their choices were determined primarily by lag rather than point value. The particular strategy adjustments we observed suggest that after experiencing List 1, participants may have concluded that, with only 12 items on a list, they could maximize performance by ignoring point value and applying the more effective strategy to all items involving the same spacing interval. When spacing was short in the later lists, participants exhibited an equally strong preference for testing over restudying with both high-value and low-value items. When spacing was long, participants chose testing more and restudying less for high-value pairs than they had on List 1, thereby choosing testing and restudying about equally often on subsequent lists, regardless of point value.

Learners' strategy choices also were affected by item difficulty, but their preference to test easy items and to restudy hard items was not altered by point value. This suggests that spacing and item difficulty may affect strategy choices through somewhat different processes, an idea to which we will return later. In any event, these findings are consistent with the general hypothesis that learners adapt their strategy choices in accordance with the prevailing learning conditions, but they provide no encouragement for the risky strategy hypothesis according to which learners choose testing for its indirect or secondary effects (e.g., diagnostic information) rather than for its direct effect on learning.

General discussion

Learners often seem to underuse retrieval practice when they choose their own learning strategies, especially when practice tests are not followed by corrective feedback. The conventional explanation is that learners do not appreciate the superior learning effects produced by retrieval practice relative to restudying. We considered an alternative hypothesis, namely, that learners may have some appreciation of the potent learning effects of retrieval practice but also believe that successful retrieval is necessary to reap its benefits. Thus, when practice tests are not followed by feedback, the choice of learning strategy would involve something analogous to an implicit risk-benefit analysis. Retrieval practice may offer greater learning benefits but entails the risk of retrieval failure and the loss of a learning opportunity. Restudying may offer a less beneficial learning experience, but it entails no risk of a lost study opportunity.

Our results were consistent with this hypothesis. Retrieval practice generally was selected to the extent that successful retrieval on the practice test was more likely. The relative tendency of learners to choose retrieval practice increased when items were easy and when the lag between the initial study opportunity and the practice test was short. It decreased when items were hard and the lag was long. Thus, there were conditions (e.g., easy items involving a short lag) in which learners clearly preferred retrieval practice over restudying, while there also were conditions (e.g., hard items involving a long lag) in which they clearly preferred restudying over retrieval practice.

Although finding a preference for retrieval practice over restudying when practice tests are not followed by feedback is consistent with the hypothesis that learners have at least some appreciation of the direct benefits of retrieval practice, it is not conclusive. Taking practice tests may be attractive for other reasons, such as the fact that attempted retrieval provides diagnostic information about how well an item has been learned even when no formal feedback is provided. The drawback to choosing testing without feedback for reasons that do not directly enhance learning, however, is that it entails the risk of completely losing an opportunity to learn if retrieval fails. In Experiment 2, we varied the point value of items with the underlying assumption that participants would be motivated to avoid risky strategies for high-value items and to apply the strategy that they believed would be more effective in fostering learning and retention. The important results are that, in studying the first list, our participants' relative preferences for testing with a short lag and for restudying with a long lag were significantly greater for high-value than for low-value items. These findings support the conclusion that, at the outset of our experiment, our participants believed retrieval practice to be a more potent learning strategy than restudying when the lag was short, although they also apparently believed that

restudying was the superior strategy when the lag was long. This conclusion is not undermined by the fact that point value was no longer effective in moderating participants' strategies after the first list. As we discussed earlier, point value is an instructional variable intended to affect participants' expectations about how to maximize final recall scores. It had the expected effect when participants had no direct experience with the actual task. However, once they had first-hand knowledge of the task, their experiences likely dominated their expectations about strategy effectiveness, reducing the influence of the instructional variable.

Our results suggest that learners' metacognitive knowledge of the benefits of retrieval practice may be more sophisticated than is often thought. Nevertheless, they often make metacognitive errors by underusing retrieval practice (e.g., Kornell & Son, 2009). The question is why. From our perspective, learners may appreciate the potential superiority of retrieval practice but also may recognize the importance of successful retrieval and, thus, the risk of retrieval failure. A sufficient explanation for the nonoptimal use of retrieval practice is that learners may be risk averse, choosing to take practice tests only when their confidence in successful retrieval is unnecessarily high. More generally, the underuse of a testing strategy may indicate a metacognitive error in which learners fail to optimally balance the risks of retrieval failure with the potential benefits of retrieval practice.

Our findings indicate that choosing a strategy in self-regulated learning involves a complex set of metacognitive judgments about the likelihood of successful retrieval, the potential benefit of each strategy for learning, and the balance between the two. The bases for judgments of retrievability were not directly addressed by our experiments but seem likely to have differed for judgments based on item difficulty and lag, respectively. Item difficulty may support relatively automatic mnemonic-based judgments (Koriat & Bjork, 2006) that rely on experiential cues such as processing fluency, in which associatively related pairs are processed more easily and therefore are thought to be easier to remember (e.g., Benjamin & Bjork, 1996; Koriat, 1997). Lag may support a more conscious, intentional, theory-based judgment in which retrievability is inferred from a belief that memory will be better after a short than after a long interval. The fact that the instructional variable, point value, did not alter the effect of item difficulty on strategy choices in Experiment 2 is consistent with a relatively automatic underlying process. However, the fact that the instructional variable did modify the effect of lag on strategy choices suggests the involvement of a more deliberate, controlled process.

Learners appear to believe that retrieval practice is a more effective learning strategy than restudying under certain circumstances, but it seems unlikely that they have a well-developed, global theory of the superiority of retrieval practice. If they did, they should recognize the superior

effects of testing in monitoring studies, but they typically do not (Tullis et al., 2013). They also might be expected to endorse retrieval practice as a learning technique in surveys, but they tend not to (e.g., Hartwig & Dunlosky, 2012; Kornell & Bjork, 2007). It seems more likely that learners hold a naïve, intuitive theory about learning that is not specific to choosing between retrieval practice and restudying but can be used for that purpose. Perhaps learners hold a naïve version of the hypothesis that more effortful processing is more effective (e.g., Carpenter & DeLosh, 2006; Pyc & Rawson, 2009), choosing retrieval practice to the extent that they intuit successful retrieval to be more effortful than restudying. Alternatively, learners might hold a naïve version of the transfer-appropriate-processing hypothesis (Roediger & Karpicke, 2006b), in which they expect practice to be more effective to the extent that it approximates the criterion task. Sorting out the possibilities is beyond the scope of the present experiments and will have to await further research.

Although we were unable to analyze or clearly interpret the recall data, our two experiments yielded parallel overall trends in recall that may be informative. Recall improved as items became easier, whether item difficulty was measured by JOLs or by normative difficulty. Consistent with the benefit of spaced practice, recall was better following long spacing than short spacing, especially when items were restudied. When items were tested, the benefit of longer spacing may have been attenuated by the tendency to test easier items. Consistent with the testing effect (e.g., Roediger & Karpicke, 2006a), testing led to better recall overall than restudying, but, again, this may partly reflect the tendency for participants to test easier items. Finally, recall of items that participants had chosen not to repractice ("done" choices) tended to suffer from the lack of additional practice. This is consistent with previous observations that learners discontinue practice prematurely to the detriment of their subsequent retention (e.g., Karpicke, 2009).

Author note Thomas C. Toppino, Department of Psychological and Brain Sciences, Villanova University; Melissa H. LaVan, Department of Psychological and Brain Sciences, Villanova University; Ryan Iaconelli, Department of Psychological and Brain Sciences, Villanova University.

Melissa H. LaVan is now a clinical psychologist in Las Vegas, NV. Ryan Iaconelli is now at the Department of Educational Studies, The Ohio State University.

References

- Abott, E. E. (1909). On the analysis of the factor of recall in the learning process. *The Psychological Review: Monograph Supplements*, 11(1), 159–177. <https://doi.org/10.1037/h0093018>
- Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General*, 138(3), 432–447. <https://doi.org/10.1037/a0015928>

- Ariel, R., Dunlosky, J., & Toppino, T. C. (2014). Contribution of degraded perception and insufficient encoding to decisions to mass or space study. *Experimental Psychology*, 61(2), 110–117. <https://doi.org/10.1027/1618-3169/a000230>
- Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(3), 940–945. <https://doi.org/10.1037/a0029199>
- Baddeley, A., & Longman, D. (1978). The influence of length and frequency of training session on the rate of learning to type. *Ergonomics*, 21(8), 627–635. <https://doi.org/10.1080/00140137808931764>
- Benjamin, A. S., & Bird, R. D. (2006). Metacognitive control of the spacing of study repetitions. *Journal of Memory and Language*, 55(1), 126–137. <https://doi.org/10.1016/j.jml.2006.02.003>
- Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. M. Reder (Ed.), *Implicit memory and metacognition* (pp. 309–338). Mahwah, NJ: Erlbaum.
- Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition* (pp. 122–144). New York, NY: John Wiley.
- Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(6), 1563–1569. <https://doi.org/10.1037/a0017021>
- 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(2), 268–276. <https://doi.org/10.3758/BF03193405>
- Dunlosky, J., & Matvey, G. (2001). Empirical analysis of the intrinsic–extrinsic distinction of judgments of learning (JOLs): Effects of relatedness and serial position on JOLs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(5), 1180. <https://doi.org/10.1037/0278-7393.27.5.1180>
- Gates, A. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, 6(40), 1–104.
- Glover, J. A. (1989). The “testing” phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, 81(3), 392–399. <https://doi.org/10.1037/0022-0663.81.3.392>
- Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, 19(1), 126–134. <https://doi.org/10.3758/s13423-011-0181-y>
- Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(1), 290–296. <https://doi.org/10.1037/a0028468>
- Izawa, C. (1966). Reinforcement-test sequences in paired-associate learning. *Psychological Reports*, 18(3), 879–919. <https://doi.org/10.1037/h0027905>
- Kang, S. H., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, 19(4/5), 528–558. <https://doi.org/10.1080/09541440601056620>
- Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, 138(4), 469. <https://doi.org/10.1037/a0017341>
- Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, 17(4), 471–479. <https://doi.org/10.1080/09658210802647009>
- Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, 319(5865), 966–968. <https://doi.org/10.1126/science.1152408>
- Koriat, A. (1997). Monitoring one’s own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126(4), 349. <https://doi.org/10.1037/0096-3445.126.4.349>
- Koriat, A., & Bjork, R. A. (2006). Mending metacognitive illusions: A comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(5), 1133. <https://doi.org/10.1037/0278-7393.32.5.1133>
- Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, 14(2), 219–224. <https://doi.org/10.3758/BF03194055>
- Kornell, N., & Son, L. K. (2009). Learners’ choices and beliefs about self-testing. *Memory*, 17(5), 493–501. <https://doi.org/10.1080/09658210902832915>
- Kuo, T.-M., & Hirshman, E. (1996). Investigations of the testing effect. *The American Journal of Psychology*, 109(3), 451–464. <https://doi.org/10.2307/1423016>
- Little, J. L., & Bjork, E. L. (2012). The persisting benefits of using multiple-choice tests as learning events. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 683–688). Austin, TX: Cognitive Science Society.
- Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, 23, 1337–1344.
- Logan, J. M., Castel, A. D., Haber, S., & Viehman, E. J. (2012). Metacognition and the spacing effect: The role of repetition, feedback, and instruction on judgments of learning for massed and spaced rehearsal. *Metacognition and Learning*, 7(3), 175–195. <https://doi.org/10.1007/s11409-012-9090-3>
- Maddox, G. B., Balota, D. A., Coane, J. H., & Duchek, J. M. (2011). The role of forgetting rate in producing a benefit of expanded over equal spaced retrieval in young and older adults. *Psychology and Aging*, 26(3), 661. <https://doi.org/10.1037/a0022942>
- McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L., III. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, 103(2), 399. <https://doi.org/10.1037/a0021782>
- McDermott, K. B., Agarwal, P. K., D’Antonio, L., Roediger, H. L., III, & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, 20(1), 3. <https://doi.org/10.1037/xap0000004>
- Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3), 402–407. <https://doi.org/10.3758/BF03195588>
- Pyc, M. A., & Dunlosky, J. (2010). Toward an understanding of students’ allocation of study time: Why do they decide to mass or space their practice? *Memory & Cognition*, 38(4), 431–440. <https://doi.org/10.3758/MC.38.4.431>
- 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(4), 437–447. <https://doi.org/10.1016/j.jml.2009.01.004>
- Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis. *Memory & Cognition*, 43(4), 619–633. <https://doi.org/10.3758/s13421-014-0477-z>
- Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15(1), 20–27. <https://doi.org/10.1016/j.tics.2010.09.003>
- Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice.

- Perspectives on Psychological Science*, 1(3), 181–210. <https://doi.org/10.1111/j.1745-6916.2006.00012.x>
- Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning—Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>
- Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, 140(6), 1432–1463. <https://doi.org/10.1037/a0037559>
- Rowland, C. A., & DeLosh, E. L. (2015). Mnemonic benefits of retrieval practice at short retention intervals. *Memory*, 23(3), 403–419. <https://doi.org/10.1080/09658211.2014.889710>
- Son, L. K. (2004). Spacing one's study: Evidence for a metacognitive control strategy. *Journal of Experimental Psychology—Learning Memory and Cognition*, 30(3), 601–604. <https://doi.org/10.1037/0278-7393.30.3.601>
- Storm, B. C., Bjork, R. A., & Storm, J. C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. *Memory & Cognition*, 38(2), 244–253. <https://doi.org/10.3758/MC.38.2.244>
- Storm, B. C., Friedman, M. C., Murayama, K., & Bjork, R. A. (2014). On the transfer of prior tests or study events to subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(1), 115. <https://doi.org/10.1037/a0034252>
- Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval questions and answers. *Experimental Psychology*, 56(4), 252–257. <https://doi.org/10.1027/1618-3169.56.4.252>
- Toppino, T. C., & Cohen, M. S. (2010). Metacognitive control and spaced practice: Clarifying what people do and why. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(6), 1480. <https://doi.org/10.1037/a0020949>
- Toppino, T. C., Cohen, M. S., Davis, M. L., & Moors, A. C. (2009). Metacognitive control over the distribution of practice: When is spacing preferred? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(5), 1352. <https://doi.org/10.1037/a0016371>
- Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition*, 41(3), 429–442. <https://doi.org/10.3758/s13421-012-0274-5>
- Wheeler, M., Ewers, M., & Buonanno, J. (2003). Different rates of forgetting following study versus test trials. *Memory*, 11(6), 571–580. <https://doi.org/10.1080/09658210244000414>
- Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory*, 20, 568–579.
- Yan, V. X., Thai, K., & Bjork, R. A. (2014). Habits and beliefs that guide self-regulated learning: Do they vary with mindset? *Journal of Applied Research in Memory and Cognition*, 3, 140–152.
- Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society*, 15(1), 41–44. <https://doi.org/10.3758/BF03329756>