MEMORY, 2017
VOL. 25, NO. 6, 764–771
http://dx.doi.org/10.1080/09658211.2016.1220579

Routledge
Taylor & Francis Group

Benefits from retrieval practice are greater for students with lower working memory capacity

Pooja K. Agarwal[a], Jason R. Finley[b], Nathan S. Rose[c] and Henry L. RoedigerIII[a]

[a]Department of Psychology, Washington University in St. Louis, St. Louis, MO, USA; [b]Department of Psychology, Fontbonne University, Clayton, MO, USA; [c]Department of Psychology, University of Notre Dame, Notre Dame, IN, USA

ABSTRACT

We examined the effects of retrieval practice for students who varied in working memory capacity as a function of the lag between study of material and its initial test, whether or not feedback was given after the test, and the retention interval of the final test. We sought to determine whether a blend of these conditions exists that maximises benefits from retrieval practice for lower and higher working memory capacity students. College students learned general knowledge facts and then restudied the facts or were tested on them (with or without feedback) at lags of 0–9 intervening items. Final cued recall performance was better for tested items than for restudied items after both 10 minutes and 2 days, particularly for longer study–test lags. Furthermore, on the 2-day delayed test the benefits from retrieval practice with feedback were significantly greater for students with lower working memory capacity than for students with higher working memory capacity ($r = -.42$). Retrieval practice may be an especially effective learning strategy for lower ability students.

ARTICLE HISTORY
Received 2 April 2016
Accepted 29 July 2016

KEYWORDS
Testing effect; retrieval practice; working memory; feedback; lag

Testing is a powerful technique to enhance learning, because the act of retrieving information from memory promotes the ability to recall material again in the future (Carpenter & DeLosh, 2005; Carrier & Pashler, 1992; see Roediger & Karpicke, 2006a, for a review). The use of retrieval practice as a learning strategy, by teachers and students, has been shown to increase students' long-term retention and transfer of knowledge to new situations (Agarwal, Bain, & Chamberlain, 2012; Butler, 2010).

In laboratory and classroom settings, several factors modulate benefits from retrieval practice, also referred to as the "testing effect" (for a review, see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). These factors include the time elapsed or the number of items between initial study and retrieval attempts (i.e., lag), the delay between initial retrieval practice and the final test (i.e., retention interval), and the presence or absence of feedback during initial retrieval. First, regarding lag, in general longer intervals between study of material and a test lead to better long-term retention, though the precise benefit from various schedules is complex and under debate (e.g., Balota, Duchek, & Logan, 2007; Karpicke & Roediger, 2007; Pyc & Rawson, 2007; Roediger & Karpicke, 2011). Second, regarding retention interval, a tradeoff is often found such that restudying improves retention in the short-term, but retrieval practice benefits learning in the long-term (e.g., Roediger & Karpicke, 2006b). In addition, shorter lags between study and retrieval trials often produce superior performance at short retention intervals, but longer lags produce superior performance at long retention intervals (Karpicke & Roediger, 2007; Whitten & Bjork, 1977). Third, benefits from retrieval practice substantially increase when feedback is provided, compared to retrieval without feedback; however, the timing of feedback following retrieval (immediate vs. delayed) and the length of the retention interval (e.g., one day vs. one week) influence its potency (Butler, Karpicke, & Roediger, 2007). In summary, lag, retention interval, and feedback all modulate benefits from retrieval practice, and various combinations of these factors produce varying degrees of enhanced learning.

Individual differences may also influence retrieval-enhanced learning (Unsworth & Engle, 2007). For instance, recent examinations reveal relationships between individual differences and retrieval difficulty (Bui, Maddox, & Balota, 2013), accessibility of retrieval cues (Unsworth, Spillers, & Brewer, 2012), and presentation duration (Unsworth, 2016). Regarding the testing effect, Wiklund-Hörnqvist, Jonsson, and Nyberg (2014) concluded that retrieval practice benefits did not differ as a function of working memory; however, their design manipulated trial type (study–study vs. study–test) between-subjects, so it cannot be determined the extent to which individual subjects exhibited the retrieval practice effect, making the null result difficult to interpret.

In a paired associate paradigm, Brewer and Unsworth (2012) found a small benefit of retrieval practice, which was significantly correlated with some individual difference

CONTACT Pooja K. Agarwal pooja@poojaagarwal.com, www.poojaagarwal.com Washington University in St. Louis, St. Louis, MO 63130, USA

measures (e.g., episodic memory) but not others (e.g., working memory). As Brewer and Unsworth noted, the relatively small testing effect they found is inconsistent with those of larger magnitude typically seen in the literature, leaving open the question whether there are "aptitude × treatment interactions" (pp. 414–415). In other words, individual benefits from retrieval may vary depending on factors known to modulate the testing effect, including lag, retention interval, and feedback.

In a follow-up study, Pan, Pashler, Potter, and Rickard (2015) conducted a replication attempt using Brewer and Unsworth's materials and general procedures. Across two experiments, one online and one in the laboratory, Pan et al. found substantial testing effects (larger than in Brewer and Unsworth), but no significant correlations between an individual difference measure (episodic memory) and benefits from testing. Pan et al. speculated that subtle procedural distinctions might have contributed to the discrepancy between the two studies. Namely, the differences in counterbalancing and the blocking or mixing of presentations may account for the increased testing effect and/or the lack of a correlation with the individual difference measure in the Pan et al. study.

Lastly, in a foreign language vocabulary paradigm, Tse and Pu (2012) found a small benefit of retrieval practice, albeit significantly correlated with a combined working memory and test anxiety measure. Echoing the concluding remarks by others, Tse and Pu acknowledged that the unexpected small testing effect might be a result of using a short lag between items, even when employing a 7-day retention interval. In other words, ascertaining a strong relationship between the testing effect and individual differences can be challenging when using shorter lags, which are known to be less potent for learning (e.g., Dunlosky et al., 2013).

To summarise, across recent studies examining individual differences, factors known to improve test-enhanced learning (lag, retention interval, and feedback) were held constant. As a result, prior studies with small testing effects and/or small correlations with individual difference measures provide an initial glimpse into the precise relationship between retrieval practice and individual differences. Our aim was to explore both the relationship between the testing effect and individual differences, as well as the relationship between individuals and optimal retrieval conditions. We examined individual differences across various levels of lag, retention interval, and feedback, variables that are known to modulate the benefits of retrieval practice. Based on the current literature, we expected to find large benefits from retrieval when testing at longer lags, with feedback, at a delayed retention interval. We also measured working memory capacity to determine whether individuals might differ in the factors needed to provide the greatest benefit from retrieval. An ideal combination of factors may not exist for all students; rather, different combinations may prove effective for different students. This research contributes to our practical understanding about the conditions that lead to the greatest benefits of test-enhanced learning, and how these conditions might be tailored to enhance learning.

Methods

Subjects

One hundred sixty-six subjects (M age = 20.0 years, 103 female) were recruited from the Washington University in St. Louis Department of Psychology human subject pool. Subjects received either credit towards completion of a research participation requirement or cash payment ($10/hour). Data from 10 subjects were excluded from analyses because they did not follow instructions or they did not return for the second session. Thus, data are reported from 156 subjects.

We note that the 156 subjects were tested at two different time periods. The initial experiment was conducted in 2008 with 104 subjects. In 2011, we added 52 more subjects from the same pool for greater power. The design and procedures used at the two time periods were identical, and analyses reported in the results section confirmed a replication of findings between the two cohorts of subjects. Accordingly, we have collapsed the remainder of the methods and results sections across the two cohorts for maximal power and variability across individuals, unless otherwise noted.

Design

We used a 2 (Trial type: study–study, study–test) × 6 (Lag: 0, 1, 3, 5, 7, 9) × 2 (Feedback for study–test trials: present, absent) × 2(Retention interval: 10 minutes, 2 days) mixed design. Trial type and lag were manipulated within subjects, whereas feedback and retention interval were manipulated between-subjects (39 subjects per cell). A non-studied baseline condition was included such that all subjects were tested on some items only during the final test (without initially studying these items) to assess how much learning had taken place during the experimental session.

Materials

One hundred ten general knowledge questions drawn from the Nelson and Narens (1980) norms were used for this experiment. An example general knowledge question used was, "What is the city in which the Baseball Hall of Fame is located?" Based on the norms, items had a 10% average recall in college students, ranging from 0.4% to 22% recall. As noted in our results section, the average baseline (non-studied) recall for the general knowledge questions found in our study was 12%, in accordance with the Nelson and Narens norms.

Of the 110 general knowledge items, 78 were used as experimental items and 32 were used as fillers to create the list structure. Thirteen sets of six facts each, equated for probability of recall, were counterbalanced across the 13 within-subject conditions (6 study–study lags, 6 study–test lags, and a non-studied baseline condition). Of the 78 critical items, subjects were presented 36 items in the study–study condition and 36 items in the study–test condition, whereas 6 items were queried only during the final test (the non-studied baseline condition). For each study–study or study–test lag (0, 1, 3, 5, 7, and 9), subjects were presented with six items and average list position was equated across trial type and lag condition.

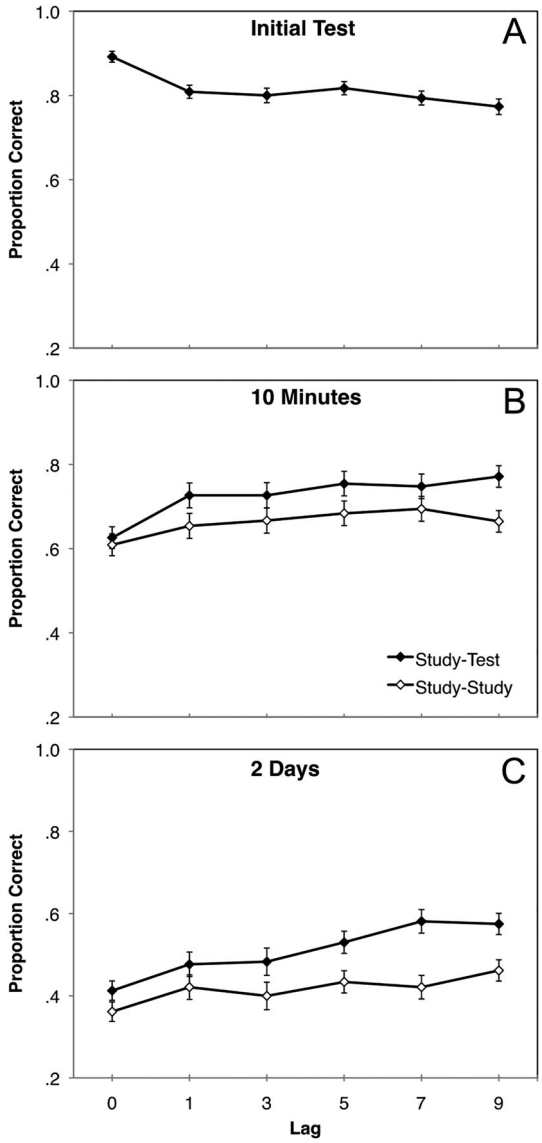

Figure 1. Mean proportion correct on initial (Panel a) and final recall tests after 10 minutes (Panel b) or 2 days (Panel c) as a function of lag and trial type, collapsed over feedback conditions. Error bars represent standard errors of the mean per lag (Panel a) or standard errors of the mean difference score per lag (Panels b and c).

Procedure

Subjects were tested individually or in small groups. They were seated at a computer and completed all learning and test phases using E-Prime 1.0 software (Schneider, Eschman, & Zuccolotto, 2002), which also provided instructions and recorded time spent on each phase of the experiment.

In the learning phase, subjects viewed 110 general knowledge questions during study and test trials. Subjects were given the following instructions:

> During study trials, you will see a trivia question with its one-word answer below it on the computer screen. Please study this pair so you can remember it later on. During test trials, you will see a trivia question with a cursor below it. Please type in the correct answer for the trivia question.

Following these instructions, subjects received one practice study–test trial (including feedback), and then moved on to the remainder of the learning phase.

For the first presentation of an item, subjects studied an intact question–answer pair for 8 seconds (e.g., What is the city in which the Baseball Hall of Fame is located? Cooperstown). For the second presentation of an item, which followed a lag of 0–9 intervening items, subjects completed either a study–study trial (for half of the items), or a study–test trial (for the other half). Study–study trials consisted of re-presentation of the intact question–answer pair for 11 seconds. Study–test trials differed by feedback condition. For the no feedback condition, subjects were shown the question and had 11 seconds to recall and type in the answer. For the feedback condition, subjects were shown the question, given 8 seconds to recall and type in the answer, and then they were shown the correct answer for three seconds. Note that total time for the second presentation of an item was equated at 11 seconds in all conditions (study–study, study–test–no feedback, and study–test–feedback).

After the learning phase, all subjects completed a working memory task on the computer for approximately 10 minutes. Specifically, subjects completed an automated operation span by Unsworth, Heitz, Schrock, and Engle (2005). Subjects were presented with a set of letters to remember, followed by a math operation to solve, followed by a recall phase in which subjects selected letters on a computer screen in the order in which the letters were presented. The span task included three sets of letters for each set size, which ranged from three to seven letters. In total, the task included 75 letters and 75 math problems. The order of set sizes was random for each participant. Unsworth et al. reported a reliability (Cronbach's α) of .78.

Subjects then received a final test either immediately following the working memory task (a 10-minute retention interval) or 2 days after the learning phase. The instructions for the final test were: "This test will look similar to the test trials earlier. You will see a trivia question at the top of the computer screen with a cursor below it. Please type in the correct one-word answer for each trivia question." During

the final test phase, subjects were presented with the 78 critical items in random order and were provided 14 seconds to type in their answer for each question.

The total time required for this procedure was approximately 90 minutes (60 min for the learning phase and working memory task, 30 min for the final test phase). Upon completion of the experiment, subjects were debriefed and thanked for their time.

Results

An alpha level of .05 was used for all tests of statistical significance except where otherwise noted. Where Mauchly's test indicated that the assumption of sphericity was violated for a within-subjects factor in an analysis of variance (ANOVA), the Greenhouse–Geisser correction was applied to the degrees of freedom. Effect sizes for comparisons of means are reported as Cohen's *d* calculated using the pooled standard deviation of the groups being compared. Effect sizes for ANOVAs are reported as $\hat{\omega}^2$ (one way) or $\hat{\omega}_p^2$ calculated using the formulae provided by Maxwell and Delaney (2004, p. 598). Standard deviations reported are uncorrected for bias (i.e., calculated using *N*, not *N* – 1).

For initial learning performance, a three-way ANOVA (cohort, lag, and feedback) showed that cohort had no significant effect and was not involved in any significant interactions ($ps \geq .245$). For final test performance, a five-way ANOVA (cohort, trial type, lag, feedback, and retention interval) showed that cohort had no significant effect and was not involved in any significant interactions ($ps \geq .136$). Furthermore, working memory capacity did not significantly differ between the two cohorts, $t(154) = 0.12$, $p = .903$. Thus, we combined the data from the two cohorts for all analyses, except where otherwise noted.

Initial learning performance

Initial learning performance is shown in Figure 1a. Reliability (Cronbach's α) was .855 for initial learning performance. Initial recall of answers to general knowledge questions declined as the lag between study and test increased. This was confirmed by a one-way ANOVA across lags, $F(5, 775) = 12.22$, $MSE = 0.261$, $p < .001$, $\hat{\omega}^2 = .030$. Follow-up *t*-tests of all 15 pairwise comparisons confirmed that lag 0 led to greater initial recall than the other lags, $ts > 5.11$, $ps < .001$, $ds > 0.41$, though differences between lags greater than 0 were not significant at the Bonferroni adjusted alpha level of .0033. We also performed an alternative analysis using regression to test the apparent decreasing pattern. For each subject, we obtained a slope using simple linear regression predicting mean initial learning performance as a function of lag. The mean slope was −.01 (SD = .02), which was significantly different from zero, $t(155) = 5.71$, $p < .001$, $d = 0.46$.

Because subjects did not receive feedback until *after* initial test trials and there was only one test per item, no effect of feedback was expected on initial learning. Accordingly, a 2 × 6 mixed ANOVA confirmed that there was neither a main effect of feedback group, $F(1, 154) = .153$, $MSE = .142$, $p = .697$, $\hat{\omega}_p^2 < .001$, nor an interaction between feedback and lag, $F(1, 154) = 1.49$, $MSE = .021$, $p = .191$, $\hat{\omega}_p^2 = .001$. Thus, the data in Figure 1a are collapsed over feedback conditions.

Final test performance

Final test performance is shown in Figure 1b (10-min retention interval) and 1c (2-day interval) as a function of whether repetitions across lags were in the study–study or study–test condition. Reliability (Cronbach's α) was .953 for final test performance. We first conducted an overall 2 × 6 × 2 × 2 mixed ANOVA (trial type × lag × feedback × retention interval) and determined that feedback (present or absent) showed no significant main effects and was not involved in any significant interactions. Thus, the data in Figure 1b and 1c and further analyses in this section were collapsed across feedback groups. Feedback may not have had an effect because performance in the tested conditions was reasonably high at the lags we used (see Figure 1a).

Second, we examined final test performance as a function of retention interval to determine if there were significant retrieval practice effects after 10 minutes and after 2 days. A 2 × 2 mixed ANOVA (trial type × retention interval) confirmed a main effect of trial type: overall final test performance was better for study–test items ($M = 62\%$, $SD = 23\%$) than for study–study items ($M = 54\%$, $SD = 23\%$), $F(1, 154) = 73.18$, $MSE = .007$, $p < .001$, $\hat{\omega}_p^2 = .036$. Forgetting occurred between 10 minutes ($M = 69\%$, $SD = 19\%$) and 2 days ($M = 46\%$, $SD = 20\%$), $F(1, 154) = 55.86$, $MSE = .074$, $p < .001$, $\hat{\omega}_p^2 = .260$. The interaction between trial type and retention interval did not reach statistical significance, $F(1, 154) = 2.67$, $MSE = .007$, $p = .104$, $\hat{\omega}_p^2 < .001$, indicating that regardless of retention interval, final performance was always greater for study–test items (10 minutes: $M = 73\%$, $SD = 19\%$; 2-day: $M = 51\%$, $SD = 21\%$) than for study–study items (10 minutes: $M = 66\%$, $SD = 20\%$; 2-day: $M = 42\%$, $SD = 19\%$). In addition, final performance for non-studied baseline items ($M = 12\%$, $SD = 14\%$) was significantly worse compared to study–study items, $t(155) = 22.89$, $p < .001$, $d = 1.48$, and study–test items, $t(155) = 27.46$, $p < .001$, $d = 1.78$, confirming that subjects were indeed learning the obscure facts and did not know most of them ahead of time.

Next, we examined final performance as a function of lag in order to determine whether there was an optimal lag for learning and whether this lag differed for the study–study and study–test conditions. Parallel analyses were conducted for the 10-min and 2-day retention interval. In both cases, the pattern in Figure 1a for initial learning was reversed at final test – whereas greater lags between initial study and restudy/test impaired performance during initial learning, they enhanced performance on the final test at both retention intervals, illustrating

the pattern Bjork (1994) described as a "desirable difficulty." The conditions leading to best initial performance led to poorest long-term retention (and vice versa).

Two separate 2 × 6 repeated measures ANOVAs (trial type × lag), one for each retention interval, confirmed main effects of lag for the 10-min retention interval, $F(5, 385) = 8.26$, MSE = .029, $p < .001$, $\hat{\omega}_p^2 = .024$, and for the 2-day retention interval $F(5, 385) = 9.94$, MSE = .036, $p < .001$, $\hat{\omega}_p^2 = .031$. Benefits from retrieval practice appeared to increase as a function of lag at both retention intervals (see Figure 1b and 1c), although the interaction between trial type and lag did not reach statistical significance at the 10-min retention interval, $F(4.5, 346.3) = 1.15$, MSE = .032, $p = .335$, $\hat{\omega}_p^2 = .001$, nor after 2 days, $F(4.4, 335.3) = 2.11$, MSE = .035 $p = .074$, $\hat{\omega}_p^2 = .003$.

Next, we performed an alternative analysis using regression to test the apparent increasing pattern. For each subject, we obtained a slope using simple linear regression predicting the mean difference score between study–study and study–test trials as a function of lag. At the 10-min retention interval, the slopes did not significantly differ from zero, $M = .006$, SD = .03, $t(77) = 1.65$, $p = .102$, $d = 0.21$. At the 2-day retention interval, however, the slopes were significantly positive, $M = .010$, SD = .03, $t(77) = 3.09$, $p = .003$, $d = 0.35$, indicating that retrieval practice benefits indeed increased as lag increased after a 2-day delay. This outcome is consistent with prior findings that testing effects often emerge on delayed tests more than on immediate tests (Roediger & Karpicke, 2006a, 2006b), and that more difficult retrieval yields greater benefits (Bjork, 1994; Finley, Benjamin, Hays, Bjork, & Kornell, 2011; Pyc & Rawson, 2009). In summary, retrieval practice improved final performance compared to restudying, both immediately (after 10 minutes) and after a delay (at 2 days); further, the benefit after a 2-day delay increased as the lag, or number of intervening items between study and retrieval trials, increased.

Associations with working memory capacity

Is there a relationship between working memory capacity and the potency of retrieval practice? To address this issue, we first examined correlations between initial and final test performance and individual differences in working memory capacity, as measured by the automatic operation span task (Unsworth et al., 2005). In keeping with Unsworth et al., we used subjects' total number of letters recalled in the correct serial position (for trials in which all letters in the sequence were correctly recalled) in the span task for all analyses. Subjects' performance on the working memory task ranged from 10 to 75 (M = 60.3, Mdn = 65.0, SD = 14.3). The maximum score for the working memory task is 75; thus, subjects in our sample demonstrated working memory capacities toward the higher end of the scale. As such, "lower" working memory in our study refers to lower task performance compared to other subjects (not low on the range of possible scores on the working memory task).

Working memory was significantly correlated with initial recall success for study–test items, $r = .31$, $t = 4.09$, $p < .001$. Next, we computed correlations between working memory scores and the difference between final performance on study–test items vs. study–study items, and did so separately for all the between-subjects conditions. These data are shown as scatterplots in Figure 2. At the 10-min retention interval (Figure 2, top panels), there was no significant correlation between working memory capacity and retrieval practice effects in the no feedback condition, $r = .18$, $t(37) = 1.09$, $p = .282$, and none in the feedback condition, $r = .11$, $t(37) = 0.69$, $p = .494$. Note that although the trend in both 10-min conditions was positive, it was not statistically significant; thus, students with differing working memory capacity benefitted equivalently from retrieval practice, either with or without feedback.

At the 2-day retention interval (Figure 2, bottom panels), there was no significant correlation in the no feedback condition, $r = -.02$, $t(37) = 0.09$, $p = .926$; however, there was a significant negative correlation in the feedback condition, $r = -.42$, $t(37) = 2.79$, $p = .008$. Note that this result replicated across our first sample ($n = 26$, $r = -.45$) and our second sample ($n = 13$, $r = -.40$), increasing our confidence in the result. Thus, for a 2-day retention interval, the lower a student's working memory capacity, the more s/he benefited from retrieval practice with feedback. We note that these specific conditions (retrieval with feedback after a 2-day delay) may be of particular relevance in applied settings, where the provision of feedback and a delay before the final test are practical and ideal for enhancing learning.

Finally, we conducted an analysis to determine whether the relationship between trial type and lag varied as a function of working memory capacity. We restrict this analysis to the 2-day retention interval group in which feedback was given during learning (Figure 2, bottom-right panel), as this is the group in which a significant correlation was observed between working memory capacity and the effect of retrieval practice. A 2 × 6 ANCOVA (trial type × lag), using working memory span as a covariate and difference scores (study–test vs. study–study) as the dependent variable, revealed no significant interactions between lag and working memory capacity, $F(5, 185) = 0.66$, MSE = .036, $p = .655$, $\hat{\omega}_p^2 < .001$, or between trial type, lag, and working memory, $F(5, 185) = 1.48$, MSE = .031, $p = .197$, $\hat{\omega}_p^2 < .001$. Follow-up t-tests at each lag showed that difference scores were greater for the lower capacity group than the higher capacity group at lags 0 and 9, $t(37) = 3.28$, $p = .002$, $d = 1.05$ and $t(37) = 3.84$, $p < .001$, $d = 1.23$, but did not significantly differ at any of the other lags (Bonferroni adjusted alpha level of .0083). Thus, although all subjects benefitted from retrieval practice, there was no obvious pattern of optimal lag between study and retrieval trials as a function of working memory capacity.

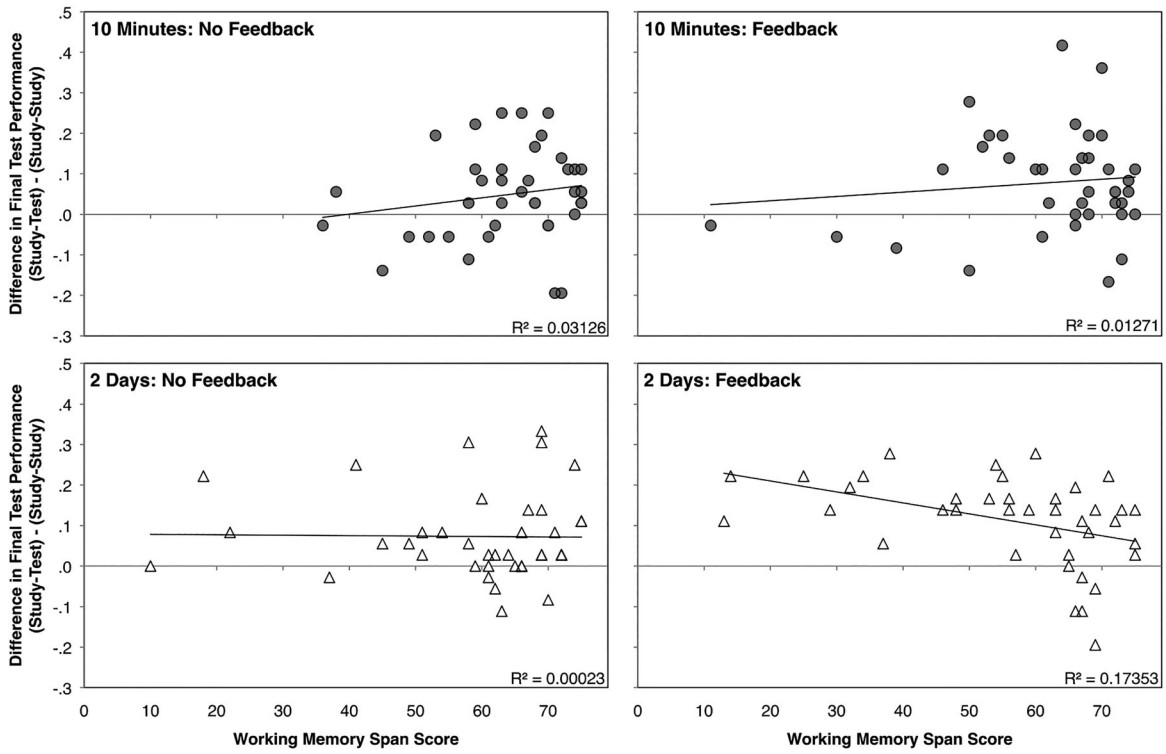

Figure 2. Difference in final test performance (study–test items minus study–study items) as a function of working memory span score, retention interval (10 minutes vs. 2 days), and feedback condition. Black lines represent the least squares linear regression.

Discussion

The primary findings from this study were: (a) retrieval practice improved performance across the board, regardless of feedback, with longer lags between study and initial test trials yielding greater benefits at the 2-day retention interval; and (b) retrieval practice with feedback yielded a greater benefit for students with lower working memory capacity at the 2-day retention interval.

We replicated the typical finding that retrieval enhances delayed performance relative to restudying (Roediger & Karpicke, 2006a) and we also confirmed previous findings that longer lags during learning enhance performance relative to shorter lags (e.g., Karpicke & Roediger, 2007; Whitten & Bjork, 1977). Even so, we were unable to determine an optimal blend of lag, retention interval, and feedback to maximise retrieval practice benefits in this paradigm. According to the mediator effectiveness hypothesis, benefits from testing are greater when the initial test is challenging because these opportunities strengthen the link between a cue and a target ("mediating information;" Carpenter, 2011; Pyc & Rawson, 2010). While it may seem counterintuitive that various combinations of "desirable difficulties" did not yield peak performance in the present study, we consider the possibility that these difficulties (retrieval, increased lag, and delayed retention interval) may have proven too challenging for students with lower working memory. In other words, at what point are difficulties for students no longer desirable? For students with lower working memory, for instance, spontaneous activation of semantic mediators, productive mediators, and/or durable mediators may fluctuate depending on the desirable difficulties present, the materials during testing, or possibly within a testing session for an individual subject. This interpretation is, of course, post hoc and needs to be examined in future research.

Surprisingly, the provision of feedback did not provide an overall additional benefit above and beyond retrieval, regardless of retention interval, possibly because performance was reasonably high on the initial test. Although we found retrieval practice with feedback improved performance at the 2-day retention interval disproportionately for lower working memory capacity students ($r = -.42$), Tse and Pu (2012) found a small benefit of testing for students with lower working memory capacity when corrective feedback was *not* provided during initial learning.

One possible explanation for this inconsistent feedback pattern relates to the bifurcation model by Kornell, Bjork, and Garcia (2011). In this framework, items that are successfully retrieved are boosted in terms of memory strength, whereas items that are not successfully retrieved nor provided feedback remain below threshold. When items are followed by feedback, however, non-retrieved items are boosted to a similar amount of memory strength as successfully retrieved items. Alternatively, this

discrepancy may be due to test-enhanced processing of feedback (e.g., Arnold & McDermott, 2013a, 2013b; Izawa, 1970; Kornell, Hays, & Bjork, 2009). Feedback allows one to identify recall errors and, thus, provides an opportunity to engage in elaborative (re)encoding of question–answer pairs in order to correct these errors on subsequent tests. Thus, it is important to bear in mind that for students with lower working memory capacity, the relationship between tests with feedback, tests without feedback, and the test–delay interaction may prove unique from students with higher working memory capacity.

We note that an appropriate examination of benefits from an intervention as a function of individual differences requires attention to several methodological issues, such as sample size and replication. While our sample size ($N = 156$) was similar to or greater than those in prior studies on retrieval practice and working memory (e.g., Brewer & Unsworth, 2012, $N = 107$; Pan et al., 2015, $N = 120$, 122; Tse & Pu, 2012, $N = 160$), future research should aim to obtain larger sample sizes. In addition, while our sample included data from two cohorts of subjects (see the "Methods" section) and we found a significant negative correlation between retrieval practice and working memory for both cohorts, additional replication is necessary to ascertain the optimal combination of lag, feedback, and retention interval for learning.

The takeaway message is that delayed benefits from testing with feedback during learning were significantly greater for students with lower working memory than for students with higher working memory capacity. This finding suggests that retrieval practice during learning, when accompanied by feedback, may serve to level the playing field for lower capacity students. Results from the present study suggest important educational implications for enhancing learning conditions for lower ability students, and further work in applied settings is necessary to sustain this conclusion.

Acknowledgements

We thank Bridgid Finn for comments on a draft of this manuscript; Andrew Butler, Jeff Karpicke, and Geoffrey Maddox for valuable discussions; and Jane McConnell for her help throughout this project.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by the National Science Foundation Graduate Research Fellowship Program and the Harry S. Truman Scholarship Foundation (awarded to the first author), and the James S. McDonnell Foundation twenty-first Century Science Initiative grant, Applying Cognitive Psychology to Enhance Educational Practice: Bridging Brain, Mind, and Behavior Collaborative Award (awarded to the fourth author).

References

Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (2012). The value of applied research: Retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review*, *24*, 437–448.

Arnold, K. M., & McDermott, K. B. (2013a). Free recall enhances subsequent learning. *Psychonomic Bulletin & Review*, *20*, 507–513.

Arnold, K. M., & McDermott, K. B. (2013b). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 940–945.

Balota, D. A., Duchek, J. M., & Logan, J. M. (2007). Is expanded retrieval practice a superior form of spaced retrieval? A critical review of the extant literature. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger, III* (pp. 83–105). New York, NY: Psychology Press.

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.

Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language*, *66*, 407–415.

Bui, D. C., Maddox, G. B., & Balota, D. A. (2013). The roles of working memory and intervening task difficulty in determining the benefits of repetition. *Psychonomic Bulletin & Review*, *20*, 341–347.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1118–1133.

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

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1547–1552.

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

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*, 633–642.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*, 4–58.

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

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology*, *83*, 340–344.

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

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

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 989–998.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Nelson, T. O., & Narens, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-

knowing ratings. *Journal of Verbal Learning and Verbal Behavior*, *19*, 338–368.

Pan, S. C., Pashler, H., Potter, Z. E., & Rickard, T. C. (2015). Testing enhances learning across a range of episodic memory abilities. *Journal of Memory and Language*, *83*, 53–61.

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

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

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*, 335.

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

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improve long-term retention. *Psychological Science*, *17*, 249–255.

Roediger, H. L., & Karpicke, J. D. (2011). Intricacies of spaced retrieval: A resolution. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: Essays in honor of Robert A. Bjork* (pp. 23–48). New York, NY: Psychology Press.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime user's guide*. Pittsburgh, PA: Psychology Software Tools.

Tse, C.-S., & Pu, X. (2012). The effectiveness of test-enhanced learning depends on trait test anxiety and working-memory capacity. *Journal of Experimental Psychology: Applied*, *18*, 253–264.

Unsworth, J. (2016). Working memory capacity and recall from long-term memory: Examining the influence of encoding strategies, study time allocation, search efficiency, and monitoring abilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 50–61.

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*, 104–132.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*, 498–505.

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2012). Working memory capacity and retrieval limitations from long-term memory: An examination of differences in accessibility. *Quarterly Journal of Experimental Psychology*, *65*, 2397–2410.

Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: The effects of spacing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 465–478.

Wiklund-Hörnqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology*, *55*, 10–16.