Journal of Memory and Language 92 (2017) 128–141

Contents lists available at ScienceDirect

Journal of Memory and Language

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

The role of mediator strength in learning from retrieval

Shana K. Carpenter [a,*], Kam Leung Yeung [b]

[a] Department of Psychology, Iowa State University, United States
[b] Department of Psychology, University of North Florida, United States

ARTICLE INFO

Article history:
Received 18 January 2016
revision received 10 June 2016
Available online 7 July 2016

Keywords:
Testing effect
Retrieval practice
Elaborative retrieval
Mediators

ABSTRACT

Previous studies have provided support for the idea that information activated during retrieval can act as a mediator that facilitates later recall of a target. Evidence for this has been obtained from a paradigm involving independent cues in which participants initially learn cue-target pairs through retrieval (*Mother:* _____) or restudying (*Mother: Child*), and later show stronger benefits of retrieval over restudy in recalling targets from final test cues that are strongly related to the original cue (*Father:* _____) compared to cues that are unrelated to the original cues (*Birth:* _____). The current study used a new paradigm to explore the role of mediators in learning from retrieval by comparing the advantage of retrieval over restudying for cue-target pairs that varied in mediator strength (i.e., the strength of the strongest first associate to the cue). Across three experiments, items higher in mediator strength (e.g., *Chalk: Crayon*, with *Chalk* producing its strongest first associate *Board* at a rate of .69) produced stronger testing effects than items lower in mediator strength (e.g., *Soup: Onion*, with *Soup* producing its strongest first associate *Chicken* at a rate of .10). Item analyses revealed that mediator strength was positively associated with final test recall of items learned through retrieval but not through restudying, and this relationship held after controlling for other linguistic properties of the cues.

Introduction

Several decades of research have shown that the process of retrieving information from memory, as compared to restudying it, produces significant enhancements in later memory for that information (for recent reviews, see Carpenter, 2012; Delaney, Verkoeijen, & Spirgel, 2010; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Kornell & Vaughn, in press; Roediger, Putnam, & Smith, 2011). The *testing effect*—also referred to as retrieval practice—has been demonstrated in hundreds of studies with a wide range of verbal materials, ranging from fairly simple word lists and word pairs (Carpenter, Pashler, Wixted, & Vul, 2008; Finn & Roediger, 2011; Kang & Pashler, 2014; Karpicke & Zaromb, 2010; Kuo & Hirshman, 1997; Peterson & Mulligan, 2013; Pyc & Rawson, 2010; Vaughn & Rawson, 2011; Vaughn, Rawson, & Pyc, 2013; Zaromb & Roediger, 2010), to more complex phrases and text passages (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Butler, 2010; Hinze & Wiley, 2011; Hinze, Wiley, & Pellegrino, 2013; Kubik, Söderlund, Nilsson, & Jönsson, 2014; Roediger & Karpicke, 2006).

The benefits of retrieval have been demonstrated in both laboratory studies (see Rowland, 2014, for a recent meta-analysis), and in classroom-based studies using realistic course materials (Carpenter, Pashler, & Cepeda, 2009; Carpenter et al., 2016; Goossens, Camp, Verkoeijen, Tabbers, & Zwaan, 2014; Jaeger, Eisenkraemer, & Stein, 2015; Karpicke, Blunt, Smith, & Karpicke, 2014; McDaniel, Anderson, Derbish, & Morrisette, 2007;

* Corresponding author at: Department of Psychology, Iowa State University, W112 Lagomarcino Hall, Ames, IA 50011, United States. Fax: +1 (515) 294 6424.
E-mail address: shacarp@iastate.edu (S.K. Carpenter).

http://dx.doi.org/10.1016/j.jml.2016.06.008
0749-596X/

McDaniel, Wildman, & Anderson, 2012). The consensus from these studies is that retrieval is a powerful mnemonic, often producing sizeable benefits on long-term memory under a variety of conditions (for recent boundary conditions to the effect, however, see Carpenter et al., 2016, and Kelly, Carpenter, & Sjolund, 2015).

Numerous empirical demonstrations of the testing effect currently outweigh our theoretical understanding of it, however. Compared to many studies that have demonstrated the benefits of retrieval, there is a shortage of studies designed to explore the more difficult question of *why* retrieval benefits memory. According to one recent account, the *elaborative retrieval hypothesis* (Carpenter, 2009, 2011), the act of retrieving a target from a cue activates cue-relevant information that becomes incorporated with the successfully-retrieved target, providing a more elaborate representation that consists of additional retrieval pathways that can be used to access the target at a later time. This activation of cue-relevant information is less likely to occur during restudy of cue-target pairs, because the target is readily available and does not need to be searched for in memory.

In one study exploring this hypothesis (Carpenter, 2011), participants learned weakly-related cue-target pairs (e.g., *Mother: Child*) by studying a list of these pairs, followed by a cued recall test (*Mother:* _____) without feedback, or a restudy opportunity over the pairs. A later final test over the same cue-target pairs demonstrated that items learned through cued recall were retained better than items learned through restudy. For different groups of participants, the final test involved cued recall of the targets from new cues that had not been shown before. Some participants received new cues that were strongly related to the original cue but not to the target. For example, according to word association norms (Nelson, McEvoy, & Schreiber, 2004), the forward associative strength between the original cue *Mother*, and the word *Father*, is .60. Other participants received new cues that were never presented and were not related to the original cue but were weakly related to the target. For example, the forward associative strength between *Mother* and *Birth* is 0, but the forward strength between *Birth* and *Child* is .02.

The purpose of providing these new cues was to evaluate whether the information activated during retrieval contributes to later retention. If the process of recalling a target from a cue (*Mother:* _____) activates cue-related information, then a word or concept closely related to the cue (*Father*) may become activated, and if the target is successfully recalled, this *mediator* could provide an additional retrieval cue to facilitate later retrieval of the target. Mediator cues like *Father* would therefore be expected to facilitate retrieval of tested items more so than restudied items, as the activation of mediators would ordinarily be less likely to occur during restudy. In contrast, a word that is unrelated to the original cue (*Birth*) would be less likely to become activated during retrieval. According to this hypothesis, the relative advantage for tested items over restudied items, therefore, would be stronger for mediator cues (*Father*) than for unrelated cues (*Birth*). This interaction was indeed observed in Carpenter's (2011, Experiment 2) study, lending support to the idea that information activated during retrieval facilitates later retrieval of target items.

One recent study replicated and extended this effect by demonstrating that the interaction reflecting the effects of elaborative retrieval—i.e., a larger advantage of testing over restudying for mediator cues compared to unrelated cues—is stronger when lists are repeated at long lags compared to short lags (Rawson, Vaughn, & Carpenter, 2015). The effects of elaboration during retrieval are expected to be stronger when items are repeated at long lags because the extended list renders the target item less accessible such that the activation of cue-relevant, non-target information during retrieval is more likely to occur. Indeed, final test performance in this study revealed the same two-way interaction reported by Carpenter (2011, Experiment 2), in addition to a three-way interaction demonstrating that the elaborative retrieval effect was stronger for items learned at long lags compared to short lags.

Previous studies have explored the role of elaboration in the testing effect by manipulating the type of cue provided during a final test—one that is either strongly associated with the original cue—i.e., a mediator—or one that is not associated with the original cue (Carpenter, 2011; Rawson et al., 2015). Important to theoretical development is the use of different methods that can measure a given construct and provide converging evidence to support it, or boundary conditions to identify its limits. To that end, the current study used a new paradigm to explore the role of mediators in learning from retrieval. Instead of manipulating the type of independent cue given on the final test, the current study used the original cues via the standard testing effect paradigm, and varied the strength of association between the cue and mediator.

Some words have stronger mediators than others. For example, according to Nelson et al.'s (2004) norms, the word *Chalk* produces its strongest forward associate *Board* at a rate of 69%. The word *Soup*, on the other hand, produces its strongest forward associate *Chicken* at a rate of only 10%. Thus, a strong forward associate—i.e., the mediator—would be more likely to be activated when *Chalk* is the cue than when *Soup* is the cue. If activation of these mediators contributes to later retrieval of the target, then the testing effect might be expected to be stronger for word pairs with stronger mediators (*Chalk – Crayon*) than for word pairs with weaker mediators (*Soup – Onion*).

To test this hypothesis, we assembled a unique set of 44 items (see Appendix) that varied in mediator strength but for which a number of other cue properties was controlled. Mediator strength was defined as the forward strength between the cue and its strongest associate, according to the norms of Nelson et al. (2004). For example, the cue *Ball* produces the mediator *Bat* 19% of the time, and the cue *Bird* produces the mediator *Fly* 21% of the time. Mediator strength across the 44 items ranged from 10% (for the cue *Soup*) to 69% (for the cue *Chalk*). For each cue, we used Nelson et al.'s norms to select a target word that was not the mediator and was weakly associated with its cue at a forward strength of about 2%, on average. Backward

associative strength between the cue and target was always 0, and none of the 44 targets ever served as the mediator for any of the other 43 cues.

A number of other cue properties was controlled to ensure that mediator strength did not co-vary with other characteristics of the cue, particularly concreteness, familiarity, imageability (obtained from the MRC Psycholinguistic Database, Wilson, 1988), and word frequency (obtained from the English Lexicon Project, Balota et al., 2007). Across the 44 items, mediator strength did not correlate with any of these other cue properties ($rs < .16$, $ps > .33$). Cues were always nouns consisting of only one syllable, and were similar to one another in length (4–6 letters).

Using these items, the current experiments utilized the standard cued-recall testing effect paradigm and explored the size of the effect as a function of mediator strength. Each of the 44 word pairs was encoded and then re-encountered as either a Test trial (involving retrieval of the target from the cue, *Ball:* _____), or a Study trial (involving re-exposure of the complete cue-target pair, *Ball: Boy*). Memory for all items was then assessed via a final test that required cued recall of each target from its original cue. If mediators play a role in learning from retrieval, then the advantage of Test over Study should be more pronounced for items that are higher in mediator strength than for items that are lower in mediator strength. Consistent with recent data on lag effects (Rawson et al., 2015), the effects of mediator strength might also be expected to be more pronounced for items that are repeatedly retrieved at longer lags compared to shorter lags. Experiment 1 was designed to provide a basic test of these hypotheses.

Experiment 1

Participants

One hundred and six students from Introductory psychology courses at Iowa State University participated to fulfill partial course requirements. The sample comprised 61% females and 39% males. Across all experiments, each participant completed the experiment individually on a personal computer. Information about participants' ages and levels of English proficiency was not collected.

Design and procedure

Each participant learned the 44 cue-target pairs during an initial learning phase, followed two days later by a final test phase. During the initial learning phase, instructions on the computer screen informed participants that they would be learning approximately 40 word pairs, and that they would later receive a memory test over those word pairs. All participants then received three encoding trials over each item, which consisted of the cue-target pair being presented by itself in the middle of the computer screen for 6 s. Following encoding, each participant re-encountered half of the cue-target pairs through retrieval (Test) and half through restudy (Study). During Test trials, the cue alone was presented (e.g., *Ball* - _____) for 6 s while participants tried to retrieve the target (e.g., *Boy*) and type it in. The program advanced after 6 s, and feedback was not provided. During Study trials, participants simply saw the cue-target pair again for 6 s.

Each participant completed the encoding phase, followed by the Test/Study phase, according to one of two schedules. For half of the participants, the 44 word pairs were divided into two separate blocks of 22 items. A unique set of 22 pairs, randomly determined for each participant, was presented for encoding three times, each time in a new random order with a 6-s presentation rate per item. Participants then saw the same list of 22 pairs again, this time with 11 pairs randomly assigned to Test and 11 to Study. The same 22 pairs were then presented two more times, each time with the same 11 pairs assigned to Test or Study. Each pair was presented for 6 s each, in a randomized order, with Test and Study trials intermixed. After completing the last Test or Study trial for the first set of 22 items, the same procedure was repeated for the other 22 items.[1]

A second group of participants learned the entire list of 44 items in a single block. Just like the other group, these participants experienced three encoding trials followed by three Study trials (SSSSSS) or three Test trials (SSSTTT), only this time the entire list of 44 items was presented three times, followed by three more presentations of the entire list with 22 items randomly assigned to Test and 22 to Study. Thus, half of the participants engaged in repeated Test vs. Study at 21-item lags (the Short Lag Group, $n = 53$), and half at 43-item lags (the Long Lag Group, $n = 53$). Given recent data showing that the use of mediators during retrieval can be moderated by lag (Rawson et al., 2015), the larger testing effect for items high in mediator strength compared to items low in mediator strength may be expected to be stronger for participants in the Long Lag Group compared to the Short Lag Group.[2]

Two days after the learning phase, participants returned to complete a final cued-recall test over all 44 items. On the final test, they were given the original cues, one at a

[1] This aspect of the design was informed by the results of a preliminary study showing that the predicted effects of mediator strength emerged when participants learned the list in 22-item blocks. In this study, 150 participants from an online sample completed one encoding trial followed by one Test or Study trial over each item, and then a final test over all 44 items after a short time interval of a few minutes. Although final test performance revealed the predicted interaction showing that retention of Test items was better for items high in mediator strength compared to items low in mediator strength (73% vs. 67%) whereas retention of Study items did not differ according to mediator strength (70% vs. 69%), $F(1,149) = 5.76$, $p = .018$, $\eta^2 = .037$, no overall advantage of Test over Study emerged, $F(1,149) = .001$, $p = .97$. Key modifications were made to this design to increase the likelihood of observing an overall testing effect, namely by delaying the final test (e.g., see Coppens, Verkoeijen, & Rikers, 2011; Roediger & Karpicke, 2006; Toppino & Cohen, 2009), and increasing the number of encoding trials to encourage successful retrieval at the time of the initial tests.

[2] Previous studies designed to explore lag effects have usually included stronger manipulations of short vs. long lags, such as 8-item vs. 35-item lags (Rawson et al., 2015), or 6-item vs. 34-item lags (Pyc & Rawson, 2009). For ease of exposition, we use the terms *short lag* and *long lag* when referring to our 21-item and 43-item lags, respectively. We note, however, that our "short lag" in the current study was longer than those typically used in previous research, and thus we might expect our lag effects to be weaker than in previous studies.

time in random order, and asked to recall and type in each target that was previously paired with each cue. The final test was self-paced, and feedback was not provided. The procedure lasted approximately 40 min during Session 1, and approximately 10 min during Session 2.

Results and discussion

Data were excluded from 8 participants who failed to return for Session 2 (5 in the Short Lag Group, and 3 in the Long Lag Group) and for 3 participants whose initial test accuracy was more than three standard deviations below the group mean. The data for the remaining 95 participants (47 in the Short Lag Group, and 48 in the Long Lag Group) were analyzed.

Initial test accuracy

Accuracy on the initial tests is shown in Table 1. A 2 × 3 (Group × Trial) Mixed ANOVA revealed a significant main effect of trial, $F(2,186) = 21.75$, $p < .001$, $\eta^2 = .19$, in that performance improved across trials. No main effect of group emerged, $F(1,93) = .18$, $p = .67$, and no interaction, $F(2,186) = .08$, $p = .92$. Inspection of initial test responses revealed that the improvement across trials was largely due to the reduction of "time-out" errors. Due to the 6-s time limit on Test trials, participants sometimes failed to respond within the allotted time, or typed in a partial response that was unidentifiable on the first trial (e.g., only the first letter of the target word), but fully correct on subsequent trials. Partial responses that could not be identified were coded as incorrect, leading to a slight underestimation of accuracy on the first trial.

Final test accuracy

On the final test, accuracy of each participants' responses was categorized according to whether participants learned items through Test or Study, and whether the item was high or low in mediator strength. High vs. low mediator strength was determined by a median split. Of the 22 items assigned to Test or Study for each participant, the 11 items with lowest mediator strength were categorized as "low," and the 11 items with highest mediator strength were categorized as "high."

Fig. 1 (upper panel) shows the mean final test scores across the four within-subjects conditions for both the Short Lag Group and the Long Lag Group. A 2 × 2 × 2 (Trial × Lag × Mediator Strength) Mixed ANOVA revealed a significant main effect of trial, $F(1,93) = 177.28$, $p < .001$, $\eta^2 = .66$, and a significant main effect of lag, $F(1,93) = 6.34$, $p = .014$, $\eta^2 = .064$. Thus, Experiment 1 demonstrated the expected benefit of Test over Study, and the benefit of long lags over short lags.

Additionally, a 3-way interaction emerged, $F(1,93) = 5.96$, $p = .016$, $\eta^2 = .06$, indicating that higher mediator strength coincided with better memory retention for Test items than for Study items, but only for the Long Lag Group and not for the Short Lag Group. The 2-way interaction between trial and mediator strength was significant for the Long Lag Group, $F(1,47) = 5.82$, $p = .02$, $\eta^2 = .11$, but not for the Short Lag Group, $F(1,46) = 1.61$, $p = .21$. In the Long Lag Group, paired-samples *t*-tests revealed a significant advantage for items that were high in mediator strength over items that were low in mediator strength, but only for Test items, $t(47) = 2.83$, $p = .007$, $d = .40$, and not for Study items, $t(47) = 1.04$, $p = .30$. The Short Lag Group demonstrated no significant advantage for items high in mediator strength over items low in mediator strength for Test items, $t(46) = .02$, $p = .98$, or for Study items, $t(46) = 1.96$, $p = .056$. No other main effects or interactions were significant, $Fs < 2.00$.

Table 1
Accuracy on initial test trials in Experiments 1 and 2.

	Trial 1	Trial 2	Trial 3
Experiment 1 (2-day delayed final test)			
Short Lag Group	.86 (.17)	.89 (.15)	.89 (.16)
Long Lag Group	.88 (.14)	.90 (.13)	.91 (.13)
Experiment 2 (1-week delayed final test)			
Short Lag Group	.88 (.15)	.92 (.13)	.92 (.14)
Long Lag Group	.88 (.13)	.90 (.14)	.90 (.12)

Note. Standard deviations are given in parentheses.

Experiment 1 provides support for the idea that mediator strength plays a role in learning from retrieval, particularly when retrieval occurs at longer lags compared to shorter lags. These findings are consistent with our predictions and with previous research showing that mediators are more likely to be activated during long-lag testing compared to short-lag testing (Rawson et al., 2015). We note, however, that final test retention for Test items was quite high in Experiment 1, and it is possible that differences between items could have been obscured by ceiling effects. During the initial tests in Session 1, participants recalled about 90% of the items correctly, on average. Performance on the final test—close to 80% in the Short Lag Group, and about 88% in the Long Lag Group—indicates that little forgetting occurred over the 2-day retention interval. Thus, Experiment 2 was designed to explore these same effects with a longer retention interval that would alleviate potential ceiling effects.

Experiment 2

Experiment 2 was a replication of Experiment 1, only this time the final test was administered after one week instead of two days. All other aspects of the procedure were identical to Experiment 1.

Participants and design

One hundred and five new students from Introductory psychology courses at Iowa State University participated to fulfill partial course requirements. The sample comprised 71% females and 29% males. Participants were randomly assigned to the Short Lag Group ($n = 50$) or to the Long Lag Group ($n = 55$), and returned after one week to complete the final test.

Results and discussion

Data were excluded from 8 participants who failed to return for Session 2 (3 from the Short Lag Group and 5 from the Long Lag Group), 2 participants who did not complete

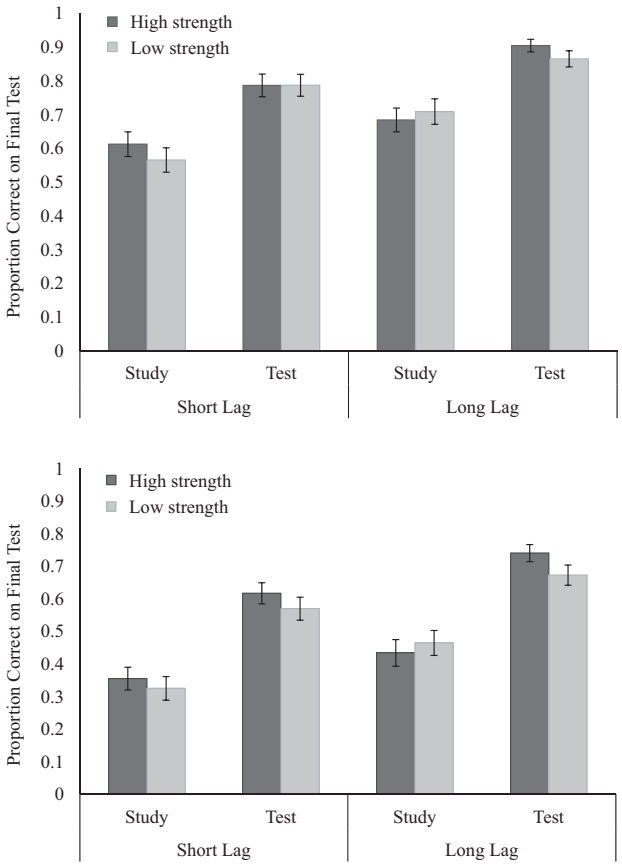

Fig. 1. Final test performance as a function of trial type, mediator strength, and lag. Experiment 1 (upper panel) assessed final test recall after a two-day delay, and Experiment 2 (lower panel) assessed final test recall after a one-week delay. In both cases, Test items that were high in mediator strength were retained better than Test items that were low in mediator strength, and this advantage was stronger at long lags compared to short lags, whereas retention of Study items did not differ as a function of mediator strength at either long or short lags.

Session 1 due to a computer crash, one participant who failed to follow instructions on the final test (i.e., typed in no responses at all), and one participant whose initial test accuracy was more than three standard deviations below the group mean. Data from the remaining 93 participants (46 in the Short Lag Group, and 47 in the Long Lag Group) were analyzed.

Initial test accuracy

Initial test recall is shown in Table 1. As in Experiment 1, a 2 × 3 Mixed ANOVA revealed a significant main effect of trial, $F(2,182) = 28.19$, $p < .001$, $\eta^2 = .24$. No effect of group emerged, $F(1,91) = .17$, $p = .68$, and no interaction, $F(2,182) = 2.35$, $p = .10$. Again, the improvement across trials appeared to be driven by a reduction in time-out errors.

Final test accuracy

Final test results closely paralleled those of Experiment 1 (see Fig. 1, lower panel). A 2 × 2 × 2 (Trial × Lag × Mediator Strength) Mixed ANOVA on final test performance revealed a main effect of trial, $F(1,91) = 231.76$, $p < .001$, $\eta^2 = .72$, and a main effect of lag, $F(1,91) = 7.04$, $p = .009$, $\eta^2 = .072$, demonstrating the predicted testing effect and lag effect. A significant main effect of mediator strength also emerged, $F(1,91) = 4.28$, $p = .04$, $\eta^2 = .045$, indicating that items high in mediator strength were retained better overall than items low in mediator strength.

A significant Trial × Mediator Strength Interaction also emerged, $F(1,91) = 4.69$, $p = .033$, $\eta^2 = .05$, indicating that the advantage for items high in mediator strength was stronger for Test items than for Study items. As in Experiment 1, this interaction was stronger for the Long Lag Group, $F(1,46) = 11.15$, $p = .002$, $\eta^2 = .20$, than for the Short Lag Group, $F(1,45) = .15$, $p = .70$, though the 3-way interaction did not reach statistical significance, $F(1,91) = 2.29$, $p = .13$, $\eta^2 = .025$. As in Experiment 1, the Long Lag Group demonstrated a significant advantage for items that were high in mediator strength over items that were low in mediator strength, but only for Test items, $t(46) = 2.75$, $p = .009$, $d = .40$, and not for Study items, $t(46) = 1.20$, $p = .24$. The Short Lag Group demonstrated no significant advantage for items high in mediator strength over items low in mediator strength for either Test items, $t(45) = 1.59$, $p = .12$, or for Study items, $t(45) = 1.04$, $p = .30$. No other effects were significant, $Fs < .60$.

Experiment 2 replicated the same pattern of results from Experiment 1, showing that mediator strength significantly coincided with better learning from retrieval than

from restudy, particularly at longer lags. This was demonstrated under conditions in which the final test was delayed by one week, reducing the likelihood that ceiling effects could be obscuring differences between conditions. Taken together, the results of Experiments 1 and 2 demonstrate that items higher in mediator strength are more likely to benefit from retrieval than from restudy.

Given the novelty of this finding and the importance of replication (Pashler & Harris, 2012; Simons, 2014), Experiment 3 was designed to provide additional data on the role of mediator strength in learning from retrieval. Alleviating ceiling effects with a one-week delayed final test, Experiment 2 demonstrated a significant interaction showing that mediator strength coincided with better memory for Test items than for Study items, and this effect was most pronounced when items were learned at long lags. The Long Lag condition is therefore of greatest interest given that the role of mediators is expected to be stronger when retrieval occurs at long lags. Thus, Experiment 3 was conducted to provide a direct replication of the Long Lag condition from Experiment 2.

Experiment 3

Participants and design

Fifty-nine new students from Introductory psychology courses at Iowa State University participated in Experiment 3 to fulfill partial course requirements. The sample comprised 76% females and 24% males. All aspects of the design were identical to the Long Lag condition of Experiment 2.

Results and discussion

Data from 8 participants were excluded from analyses. Five of these participants failed to return for Session 2, and data from 3 additional participants were lost due to a computer crash or experimenter error during Session 1. Data from the remaining 51 participants were analyzed.

Initial test accuracy

Accuracy on the initial test in Experiment 3 was 72% on Trial 1 (SD = 26%), 73% on Trial 2 (SD = 27%), and 75% on Trial 3 (SD = 26%). A one-way Repeated Measures ANOVA revealed a significant improvement in recall across trials, $F(2,100) = 11.05$, $p < .001$, $\eta^2 = .18$, which again appeared to be driven by a reduction in time-out errors.

Final test accuracy

Final test performance across the four within-subjects conditions is displayed in Table 2. A 2 × 2 (Trial × Mediator Strength) Repeated Measures ANOVA revealed a significant main effect of trial, $F(1,50) = 142.41$, $p < .001$, $\eta^2 = .74$, confirming again the advantage of Test over Study. A significant Trial × Mediator Strength Interaction also emerged, $F(1,50) = 4.10$, $p = .048$, $\eta^2 = .076$. As in the previous experiments, items high in mediator strength were retained better than items low in mediator strength, but this only occurred for Test items, $t(50) = 2.18$, $p = .03$, $d = .31$, and not for Study items, $t(50) = .83$, $p = .41$. There was no overall main effect of mediator strength, $F(1,50) = 1.01$, $p = .32$.

Table 2
Proportion of items recalled on the final test in Experiment 3 as a function of mediator strength and trial type.

	Low mediator strength	High mediator strength
Study	.31 (.25)	.29 (.26)
Test	.54 (.28)	.59 (.25)

Note. Standard deviations are given in parentheses.

Findings from Experiment 3 confirmed the same benefit of Test over Study that was observed in Experiments 1 and 2. Experiment 3 also confirmed the same interaction from Experiments 1 and 2, showing that mediator strength coincides with better memory retention for Test items compared to Study items. Consistent with Experiments 1 and 2, this interaction emerged under conditions in which items were learned at long lags.

Supplemental analyses

Continuous analyses by items

The data from all three experiments were analyzed using by-item analyses to assess the relationship between mediator strength and final test recall. Mediator strength was treated as a continuous variable and correlated with the proportion of participants who correctly recalled each of the 44 items on the final test in the Study and Test conditions. For Test items, the proportion of participants recalling each item correctly on the final test was correlated significantly with the proportion of participants recalling each item correctly on the initial test (across experiments including all five Test conditions, rs ranged from .37 to .86, ps < .015).[3] This does not seem particularly surprising, as when corrective feedback is withheld after a retrieval attempt participants would be unlikely to recall items on the final test that they could not recall initially. The influence of any variable, such as mediator strength, on final test recall would thus only occur if the items could be successfully recalled initially. In measuring the relationship between mediator strength and final test recall across Test items, therefore, we controlled for the rate of initial test recall.[4]

If mediators play a role in learning from retrieval, then a positive relationship between final test recall and mediator strength should occur, and this relationship should be stronger for Test items than for Study items. To explore

[3] The proportion of participants recalling each item correctly on the initial test in Experiment 1 ranged from 72% to 100% for the Short Lag Group, and from 78% to 100% for the Long Lag Group. The proportion recalling each item correctly in Experiment 2 ranged from 75% to 100% for the Short Lag Group, and from 76% to 100% for the Long Lag Group. Finally, in Experiment 3, the proportion ranged from 52% to 100%.

[4] Initial test recall by items was defined as the proportion of participants who recalled each item correctly at least once across the three initial test trials. Across items, initial test recall did not correlate significantly with mediator strength. This was the case in Experiment 1 for both the Short Lag Group ($r = -.10$) and the Long Lag Group ($r = .22$), in Experiment 2 for both the Short Lag Group ($r = .05$) and the Long Lag Group ($r = -.03$), and in Experiment 3 ($r = .003$), all ps > .14.

Table 3
By-item regression analyses for predicting final test recall for test items using initial test recall, mediator strength, and additional cue properties as predictors.

Predictors	Adjusted R^2	*B*	SEB	β	*p*
Experiment 1 Long Lag Test	.71				<.001
Initial Test Recall		1.07	.111	.867	<.001
Mediator Strength		.016	.051	.030	.749
Imageability		.001	.001	.079	.418
Concreteness		.000	.000	.053	.616
Familiarity		.000	.000	.001	.997
Frequency		.004	.009	.057	.681
Experiment 1 Short Lag Test	.34				<.001
Initial Test Recall		.763	.179	.538	<.001
Mediator Strength		.106	.081	.183	.197
Imageability		.001	.001	.127	.385
Concreteness		.000	.000	−.127	.416
Familiarity		−.001	.000	−.388	.058
Frequency		.016	.014	.226	.279
Experiment 2 Long Lag Test	.50				<.001
Initial Test Recall		1.10	.201	.600	<.001
Mediator Strength		.229	.108	.258	.040
Imageability		.003	.002	.265	.041
Concreteness		.000	.001	−.092	.497
Familiarity		−.001	.000	−.330	.064
Frequency		.033	.019	.305	.097
Experiment 2 Short Lag Test	.27				.007
Initial Test Recall		.970	.303	.460	.003
Mediator Strength		.083	.134	.093	.538
Imageability		.005	.002	.428	.011
Concreteness		−.001	.001	−.127	.442
Familiarity		−.001	.001	−.301	.166
Frequency		.013	.024	.119	.598
Experiment 3 Long Lag Test	.64				<.001
Initial Test Recall		.971	.145	.702	<.001
Mediator Strength		.163	.094	.181	.090
Imageability		.001	.001	.116	.307
Concreteness		.000	.001	−.104	.390
Familiarity		−.001	.000	−.197	.195
Frequency		.010	.017	.090	.563

Note. Initial test recall refers to the proportion of participants who recalled each item correctly at least once across the three test trials.

these effects, multiple regression analyses were conducted to predict final test recall for Test items after controlling for initial test recall and the additional cue properties of concreteness, familiarity, imageability, and frequency. The same analysis applied to Study items entered the four cue properties as predictors. These outcomes are reported in Table 3 for Test items, and Table 4 for Study items.

After controlling for initial test recall and the other cue properties, a unique portion of the variance in final test recall of Test items was accounted for by mediator strength in the Long Lag Test condition of Experiment 2 ($\beta = .26$, $p = .04$) and Experiment 3 ($\beta = .18$, $p = .09$). The same effect did not occur in the Long Lag Test condition of Experiment 1. However, in Experiment 1 very little forgetting occurred for Test items. When final test recall was conditionalized upon correct initial test recall, 95% of the items (42 of the 44) were recalled at a rate of .90 or higher on the final test, likely creating a truncated range of final test scores. When forgetting was induced by a one-week delay before the final test in Experiments 2 and 3, mediator strength accounted for a unique portion of the variance in final test recall, but only for Test items learned at longer lags.

The same analysis applied to Study items revealed no significant relationship between mediator strength and final test recall (see Table 4). In fact, much of the time the relationship was slightly *negative*, indicating that Study items higher in mediator strength were remembered less well than Study items lower in mediator strength—the opposite pattern to that observed for Test items.

Table 5 reports the by-item regression analysis applied to the combined data from the Long Lag Groups of Experiments 2 and 3. Mediator strength again accounted for a unique portion of the variance in final test recall of Test items ($\beta = .25$, $p = .015$), but not Study items ($\beta = -.15$, $p = .38$). This pattern is consistent with the interactions reported earlier and supports the idea that mediator strength is positively associated with learning from retrieval but not from restudy.

Error responses on the initial tests

As an additional supplemental analysis, we examined error response that occurred on the initial test trials across the three experiments. Examination of these error

Table 4
By-item regression analyses for predicting final test recall for study items using mediator strength and additional cue properties as predictors.

Predictors	Adjusted R^2	*B*	SEB	β	*p*
Experiment 1 Long Lag Study	.27				.004
Mediator Strength		−.004	.116	−.005	.975
Imageability		.005	.002	.487	.003
Concreteness		−.001	.001	−.150	.362
Familiarity		−.001	.000	−.431	.045
Frequency		.011	.021	.112	.608
Experiment 1 Short Lag Study	.12				.080
Mediator Strength		.091	.152	.096	.554
Imageability		.005	.002	.350	.042
Concreteness		−.001	.001	−.245	.178
Familiarity		−.001	.001	−.318	.172
Frequency		.008	.027	.069	.774
Experiment 2 Long Lag Study	.12				.072
Mediator Strength		−.100	.157	−.103	.527
Imageability		.004	.002	.269	.111
Concreteness		−.002	.001	−.357	.052
Familiarity		−.001	.001	−.301	.194
Frequency		.007	.028	.058	.809
Experiment 2 Short Lag Study	.12				.072
Mediator Strength		−.031	.156	−.032	.842
Imageability		.004	.002	.265	.116
Concreteness		−.002	.001	−.334	.069
Familiarity		−.001	.001	−.354	.128
Frequency		.016	.028	.134	.576
Experiment 3 Long Lag Study	−.04				.628
Mediator Strength		−.100	.174	−.101	.568
Imageability		.002	.003	.147	.420
Concreteness		−.001	.001	−.166	.396
Familiarity		−.001	.001	−.195	.436
Frequency		.021	.031	.176	.499

Table 5
By-item regression analyses for predicting final test recall for long lag test and long lag study items in Experiments 2 and 3 combined.

Predictors	Adjusted R^2	*B*	SEB	β	*p*
Long Lag Test	.68				<.001
Initial Test Recall		1.19	.161	.708	<.001
Mediator Strength		.204	.080	.247	.015
Imageability		.002	.001	.175	.096
Concreteness		.000	.000	−.066	.554
Familiarity		−.001	.000	−.274	.057
Frequency		.023	.014	.236	.112
Long Lag Study	.08				.156
Mediator Strength		−.123	.138	−.147	.379
Imageability		.003	.002	.254	.142
Concreteness		−.001	.001	−.284	.129
Familiarity		−.001	.001	−.276	.246
Frequency		.015	.025	.146	.552

Note. Initial test recall refers to the proportion of participants who recalled each item correctly at least once across the three test trials.

responses revealed that participants made errors of commission that were related to the cues. Most often, that error was the mediator. For example, when given the cue *Broom* - _____, participants mistakenly typed the word *Sweep* (the strongest forward associate of *Broom*, with mediator strength of .504) instead of the target word *Floor*. These "mediator intrusions" occurred 23% of the time in Experiment 1, 16% of the time in Experiment 2, and 15% of the time in Experiment 3.

Other times participants typed in an error response that was not the mediator, but was listed among the forward associates of the cue according to the norms of Nelson et al. (2004). For example, the cue *Broom* - _____ sometimes elicited the errors *Mop* or *Stick*, both of which are

forward associates of *Broom*. Although the mediators were the most common commission errors, other associates to the cues (excluding the mediators) comprised an additional 30% of commission errors in Experiment 1, an additional 33% in Experiment 2, and an additional 34% in Experiment 3. Thus, across the three experiments, participants' errors of commission commonly consisted of words that were connected to the cues by pre-existing associations. If these errors represent information that is activated during retrieval, then cue-relevant responses appear to be activated during the process of recalling a target from a cue, consistent with the premise of the elaborative retrieval hypothesis.

General discussion

These data support the idea that mediators activated during retrieval contribute to later retention. Experiment 1 yielded an overall benefit of Test over Study, in addition to the predicted interaction showing that items higher in mediator strength coincided with better memory retention for Test items than for Study items. As predicted, this interaction only occurred for the Long Lag Group and not for the Short Lag Group. Implementing a longer retention interval, Experiments 2 and 3 demonstrated the same interaction between testing and mediator strength for the Long Lag Group. Analyses by items revealed that final test performance was positively associated with mediator strength, particularly for Test items learned at long lags. Mediator strength did not correlate positively with final test recall of Study items in any of the experiments.

These results are consistent with those of previous studies showing that mediators play a role in learning from retrieval (e.g., Carpenter, 2011). In particular, Experiments 1 and 2 are consistent with the results of one recent study showing that the role of mediators in retrieval is stronger when repeated retrieval occurs at long lags compared to short lags (Rawson et al., 2015). These previous studies used independent cues to measure the effects of mediators. The current study used original cues via the standard testing effect paradigm, and found that cue-target pairs containing stronger mediators were more likely to benefit from retrieval than were cue-target pairs containing weaker mediators, and this was particularly true for retrieval that took place at longer lags. This finding was specific to retrieval, as learning via restudy produced no differences in final test recall as a function of mediator strength.

These results confirm the findings of many studies demonstrating benefits of retrieval over restudy (e.g., Carpenter, 2012; Delaney et al., 2010; Dunlosky et al., 2013; Roediger et al., 2011; Rowland, 2014), and benefits of longer lags over shorter lags (e.g., Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012; Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008). These variables have been known to interact as well, such that the benefits of retrieval over restudy are greater at longer lags than at shorter lags (Rawson et al., 2015; Thios & D'Agostino, 1976). This two-way interaction between testing and lag did not occur in the current study. However, this was likely due to the fact that our "short" lag was longer than the short lags that have been used in previous studies, which have typically consisted of list items separated by only 4 (Thios & D'Agostino, 1976) to 6 (Pyc & Rawson, 2009) to 8 other items (Rawson et al., 2015). Comparatively, a list of 22 items could require more effortful retrieval from long-term memory, leading to benefits of retrieval over restudying that were larger than those typically seen in previous studies. Indeed, at least one study has shown that lag effects for tested items tend to occur in a graded fashion, such that memory retention is lowest following a short lag between encoding and initial test, slightly higher following an intermediate lag, and highest following a long lag (Whitten & Bjork, 1977). Our 22-item list likely represented an intermediate lag that exceeded short-term memory. With further increases in lag—i.e., the 44-item list—the effort associated with retrieval increases (Pyc & Rawson, 2009), and in this case the effects of elaboration during retrieval may be more likely to occur, accounting for the observed effects of mediators on retrieval at longer lags in both previous research (Rawson et al., 2015) and in the current study.

The beneficial effects of mediator strength on learning from retrieval are consistent with a number of recent studies showing that retrieval conditions that encourage the activation of cue-relevant information enhance later recall of the targets that were learned from those cues. In a study by Kornell, Hays, and Bjork (2009), participants learned weakly-associated cue-target pairs by first trying to generate the target when given the cue (e.g., *Whale:* _____) and then seeing the cue and correct target together (e.g., *Whale: Mammal*), versus simply seeing the cue and correct target the whole time. Even though the generated answers were always incorrect, performance on a later test over the same cue-target pairs was significantly higher when participants generated the wrong answer and then received feedback, compared to never trying to generate the answer. Kornell et al. reasoned that the act of generating the target could activate meaningful semantic information related to the cue that becomes integrated with meaningful information about the target when it becomes available, such that the generation of this "incorrect information can serve as a mediator, connecting the question with the correct answer" (p. 996). Indeed, this idea has received support from later studies directly examining the beneficial role of mediators in the testing effect (e.g., Carpenter, 2011; Rawson et al., 2015).

Subsequent studies have replicated Kornell et al.'s (2009) findings and have found that, at the time of the final test, participants are able to successfully remember their original guesses. Furthermore, retrieval of the correct target is more likely to occur when participants are also able to retrieve their original guesses. This has been observed when the final test occurs relatively immediately after learning (e.g., Knight, Ball, Brewer, DeWitt, & Marsh, 2012; Vaughn & Rawson, 2012) and also when the final test occurs after a delay (Yan, Yu, Garcia, & Bjork, 2014; for similar findings using slightly different types of tasks, see also Butler, Fazio, & Marsh, 2011; Pyc & Rawson, 2010), adding support to the idea that these initial guesses

may serve as mediators that facilitate later recovery of the target.

Other studies using this same basic paradigm have found that the positive effects of guessing are stronger when the cue and target are semantically related. For example, Huelser and Metcalfe (2012; see also Grimaldi & Karpicke, 2012) had participants learn weakly related word pairs (with forward associative strength of 5%, on average, according to the norms of Nelson et al., 2004) vs. unrelated word pairs, and found that the benefits of generating erroneous guesses only occurred for the related word pairs. Knight et al. (2012) replicated this same result and also found that the tendency to retrieve correct targets after retrieving initial guesses occurred more often for related word pairs than for unrelated word pairs. This provides further support for the idea that erroneous guesses can act as mediators that connect the cue and target, and that the accessibility of these mediators is greater when they are generated based on a semantic association with the cue. Indeed, other studies have found that the benefits of generating erroneous guesses are eliminated under conditions where semantic processing of the cue is discouraged, such as when the cue provides a constraining stem that restricts the response options (e.g., *tide: wa__*, Grimaldi & Karpicke, 2012; see also Bridger & Mecklinger, 2014), or when the cues provided are lexically related to the targets (e.g., *ho___: house*) rather than conceptually related (*a flower: tulip*) (Cyr & Anderson, 2015).

This pattern of findings is consistent with the premise that generating information based on pre-existing associations to the cue during retrieval enhances later retrieval of the target by providing a mediator between the cue and target that can serve as a pathway for facilitating retrieval of the target at a later time. Having been generated from the cue during the search to recover the target, this mediating information would be expected to benefit later retrieval in the cue-to-target direction more so than in the target-to-cue direction. This idea has received some preliminary support from studies that have explored the directionality of the testing effect with cue-target pairs, and have observed an asymmetrical pattern favoring "forward" recall over "backward" recall (Vaughn & Rawson, 2014; see also Carpenter, Pashler, & Vul, 2006).

The activation of mediating information is assumed to occur during retrieval of a target from a cue, and thus would not be expected to occur to the same extent during mere presentation or restudy of the to-be-learned information. Elaborative activities during *encoding*, therefore, such as the keyword mnemonic (Karpicke & Smith, 2012), or free association of related words to cues or targets presented in isolation (Lehman & Karpicke, in press, Experiments 3–5; Lehman, Smith, & Karpicke, 2014), would not be expected to engage the same type of cued retrieval-based processing that is proposed by the elaborative retrieval hypothesis. In a recent study, Lehman and Karpicke (in press, Experiments 1–2) found that retrieving targets (e.g., *Child*) from word stem cues (e.g., *Mother: Ch___*) did not result in activation of the mediator "Father," as evidenced through comparable lexical decision response times to mediators (*Father*) vs. words that were unrelated to the cue (*Banquet*) following retrieval (*Mother: Ch___*) vs. restudy (*Mother: Child*). However, the word stem cue used in this particular task likely constrained responses to the target word (i.e., starting with "ch"), and discouraged activation of other words related to the cue. Consistent with the studies reviewed earlier, constrained cues are less likely than unconstrained cues to engender the activation of mediators that are believed to facilitate connections between cues and targets (Bridger & Mecklinger, 2014; Cyr & Anderson, 2015).

The current results also indicate that it is not necessarily the *quantity* of information that can be activated during retrieval, but the strength of the mediating information and how it facilitates cue-target connections. In the current materials, cue items that were higher in mediator strength necessarily contained a smaller overall number of forward associates, $r = -.72$, $p < .001$, indicating that the sheer quantity of information associated with a cue, without regard to its strength of activation or quality to serve as effective mediating information, is not the locus of the effect. As such, the idea that mediators activated during retrieval can facilitate later retention is not at odds with phenomena such as the *fan effect* (Anderson, 1974; Anderson & Reder, 1999; Cook, Marsh, Hicks, & Martin, 2006) or the *cue overload principle* (Anderson & Neely, 1996; Parkin, 1980; Watkins & Watkins, 1975), showing that a greater number of associations to a given cue can interfere with memory for any one of those associations. Whereas these effects are believed to be accounted for by interference among multiple associations, the beneficial effect of mediators is more likely attributed to the role that they play in connecting the cue and target. The extent to which cue-related information facilitates memory likely depends on the degree to which it interferes with, versus *integrates*, what is being learned. Indeed, studies have shown that the fan effect can be reduced or eliminated when multiple associations can be integrated into a unified concept (e.g., Moeser, 1979; Radvansky, Zacks, & Hasher, 1996; Smith, Adams, & Schorr, 1978), and the effect can be reversed under conditions in which the integrated associations are causally linked with one another (Myers, O'Brien, Balota, & Toyofuku, 1984).

Future research would benefit from further explorations of how mediators are used during retrieval, and whether different strategies exist for using mediators that may yield different effects on long-term retention. In addition to the possibility that mediating information activated from a cue becomes directly connected with the target, a non-mutually-exclusive possibility is that mediating information serves as an effective retrieval cue by reminding participants about what the answer is *not*. At the time of a final test, participants may see a cue (e.g., *Night: _____*) that activates a strongly associated concept (*Day*), but they remember retrieving this item and that this strong associate was not the right answer, leading them to eventually recall the correct answer *Train*. Thus, if mediators are activated during retrieval and the correct target is successfully retrieved, the retrieval process could serve to effectively distinguish the mediator from the correct answer. Such a possibility is consistent with the finding that successful recall of targets on a final test coincides with successful recall of initial erroneous guesses (Butler et al., 2011; Cyr

& Anderson, 2015; Knight et al., 2012; Vaughn & Rawson, 2012; Yan et al., 2014) and with successful recall of previously-generated mediators designed to link a cue and target (Pyc & Rawson, 2010). A worthwhile question to explore in future research is whether mediating information facilitates retrieval of target information across a range of time intervals, or whether "mediator intrusions" return after a long enough delay. The relationship between mediator strength and later recall of tested items may not be linear. After enough time has passed, the memory strength of a strong mediator could exceed that of the correct answer, such that items with the strongest mediators may not actually be the best retained in the long run.

The findings of the current study can be considered in light of additional explanations that have been offered to explain the benefits of retrieval. These accounts have pointed to the importance of the type of processing and retrieval success at the time of initial test, the processing required at the time of the final test, and the role of contextual features as key contributors. According to the *bifurcated distribution model* (Kornell, Bjork, & Garcia, 2011; see also Halamish & Bjork, 2011), the long-term advantage in memory for tested items over restudied items depends upon retrieval success at the time of the initial test—items successfully retrieved receive a strong boost in memory strength, those not retrieved receive no boost, and those that are restudied receive an intermediate boost—and also the threshold for retrieval at the time of the final test, such that advantages of tested items are more likely to emerge when the threshold is higher, as in the case of a delayed final test rather than an immediate final test. The *episodic context account* (Karpicke, Lehman, & Aue, 2014) proposes that items are initially encoded along with contextual information associated with them, and that during retrieval participants attempt to reinstate the context in which the items were encoded. Successful retrieval results in an updating of contextual representations that include both the original and current context, such that the retrieved items "become associated with a variety of contextual features that serve as effective retrieval cues on later tests" (Karpicke et al., 2014, p. 259).

The idea that mediators contribute to the benefits of retrieval is not mutually exclusive from these accounts, and can further their specificity by proposing an explanation for why some items may be more likely than others to benefit from retrieval. For example, the increment in memory strength that results from retrieval, as conceptualized by the bifurcation model, could be influenced in part by the presence of mediating information, such that item properties or retrieval situations that encourage the use of mediators could lead to a greater boost in memory strength as a result of retrieval. A straightforward premise of the episodic context account is that the retrieval process involves attempts to reactivate information previously encountered. The elements of the retrieval context that have become activated or strengthened as a result of this process can then serve as effective retrieval cues, and in this regard, the general process described by the episodic context account does not appear to differ fundamentally from that proposed by the elaborative retrieval account. Particularly with regard to cued retrieval tasks of the type explored here, both accounts propose that information that is based on the encoded stimuli (i.e., cue-target pairs) is activated, reinstated, brought to mind, etc. during retrieval, and this information (i.e., mediators, contextual elements, etc.) can serve as effective cues for retrieval of the stimuli again later. By this logic, mediators as conceptualized by the elaborative retrieval account and elsewhere (e.g., Cyr & Anderson, 2015; Knight et al., 2012; Kornell et al., 2009; Pyc & Rawson, 2010) could be one type of "contextual element" that is activated or reinstated during retrieval and that can promote later memory for the retrieved information.

The factors underlying the testing effect are likely to be multifaceted, such that the effect is best accounted for by a number of potentially inter-related mechanisms. The activation of mediators during retrieval does not account for all instances of the effect. In the current study we observed large benefits of retrieval over restudy even for items that were low in mediator strength, and even for items that were learned at short lags. Further, it is likely that factors other than, or in addition to, activation of mediators underlie the benefits of testing that have been observed with materials that do not encourage obvious semantic or conceptual associations between items, such as learning to connect unfamiliar symbols to words (e.g., Coppens et al., 2011; Kang, 2010), or learning a spatial representation of locations within an environment (Kelly et al., 2015; Rohrer, Taylor, & Sholar, 2010). There is also evidence that the time course of mediators may be limited, in that they are activated during early stages of learning but cease to be utilized after a cue-target pair has become well-learned (e.g., Kole & Healy, 2013).

In conclusion, these data provide converging evidence for the role of mediators in learning from retrieval by showing that items higher in mediator strength are more likely to benefit from retrieval than items lower in mediator strength, and that this benefit is most likely to occur for items retrieved at longer lags. These findings help advance our theoretical understanding of the testing effect by pointing to the activation of mediators during retrieval as one mechanism that contributes to the benefits of testing over restudying. Future research is encouraged that can shed light on additional mechanisms that may account for the benefits of retrieval in a broad range of tasks.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant DUE-1504480. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank Rosa Cai, Brianna Culich, Chad Fernandez, Anne Herlache, Torie Homervold, Shuhebur Rahman, Vivian Spilde, Rachel Teague, and Emma van Buer for their assistance with data collection.

Appendix. Items used in the current study

Cue	Target	Cue-target strength	Mediator	Mediator strength	Other cue properties			
					Concreteness (MRC)	Familiarity (MRC)	Imageability (MRC)	Frequency (HAL_Log)
Ball	Boy	.011	Bat	.190	615	575	622	10.558
Bird	Tree	.020	Fly	.212	602	592	614	9.856
Blood	Skin	.011	Red	.337	613	571	620	10.855
Bomb	Fire	.013	Explode	.150	595	566	606	9.641
Boot	Army	.020	Shoes	.338	595	566	604	10.689
Bread	Meat	.013	Butter	.487	622	611	619	9.112
Broom	Floor	.016	Sweep	.504	613	547	608	6.397
Chair	Bed	.013	Table	.314	606	617	610	9.830
Chalk	Crayon	.020	Board	.694	634	560	601	7.416
Child	Girl	.020	Baby	.202	581	585	619	11.078
Claw	Tiger	.020	Cat	.177	587	445	600	7.693
Clock	Radio	.012	Time	.652	591	608	614	10.106
Coin	Bill	.012	Money	.436	581	564	603	8.76
Deer	Woods	.014	Doe	.132	631	509	624	8.499
Feast	Party	.018	Food	.248	542	457	610	7.555
Fence	Cage	.024	Picket	.150	597	526	611	8.571
Geese	Eggs	.020	Ducks	.291	597	421	603	6.461
Golf	Grass	.014	Ball, Club[a]	.155	540	503	616	9.254
Grave	Dirt	.022	Death	.244	535	501	619	8.926
Harp	Flute	.012	Music	.412	591	430	621	7.419
Heart	Body	.016	Beat	.183	605	578	617	10.732
Horse	Dog	.012	Ride	.261	613	560	624	10.080
Jail	Thief	.013	Prison	.261	590	539	608	9.176
Kite	Paper	.017	Fly	.575	592	481	624	7.76
Knife	Gun	.013	Fork	.327	612	573	633	8.871
Lamb	Wolf	.016	Sheep	.333	633	519	614	8.144
Lunch	Pail	.013	Dinner	.269	552	616	602	9.314
Milk	Water	.024	Cow	.388	670	588	638	9.629
Moose	Bull	.018	Animal	.182	616	518	604	8.153
Mouth	Face	.022	Wash	.201	568	572	613	10.439
Neck	Bone	.018	Head	.333	587	576	622	9.605
Night	Train	.019	Day	.686	496	636	607	11.488
Nurse	Needle	.014	Doctor	.547	588	537	617	8.801
Prince	Crown	.020	Princess	.409	542	506	606	9.31
Road	Sign	.020	Street	.348	583	604	609	11.108
Rock	Mountain	.019	Stone	.269	600	583	612	10.698
Roof	Rain	.016	House	.307	586	552	604	8.635
Rose	Bud	.021	Flower	.363	608	556	623	9.869
Shore	Waves	.014	Beach	.459	574	531	624	8.718
Skunk	Stripe	.016	Smell	.559	648	519	652	6.786
Snake	Spider	.012	Bite	.279	621	501	627	8.560
Soap	Cloth	.011	Clean	.239	598	594	600	8.992
Soup	Onion	.024	Chicken, Tomato[a]	.103	615	576	604	8.707
Truck	Bus	.014	Car	.264	595	620	621	9.230

[a] Two of the cues (Golf and Soup) each produced two first associates of equal strength.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., III, & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22*, 861–876.

Anderson, J. R. (1974). Retrieval of proposition information from long-term memory. *Cognitive Psychology, 6*, 451–474.

Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General, 128*, 186–197.

Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memory retrieval. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (pp. 237–313). San Diego, CA: Academic Press.

Balota, D. A., Yap, M. J., Cortese, M., Hutchison, K. A., Kessler, B., Loftis, B., ... Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*, 445–459.

Bridger, E. K., & Mecklinger, A. (2014). Errorful and errorless learning: The impact of cue-target constraint in learning from errors. *Memory & Cognition, 42*, 898–911.

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., Fazio, L. F., & Marsh, E. J. (2011). The hypercorrection effect persists over a week, but high confidence errors return. *Psychonomic Bulletin & Review, 18*, 1238–1244.

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

Carpenter, S. K. (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. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*, 279–283.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review, 24*, 369–378.

Carpenter, S. K., Lund, T. J. S., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review, 28*, 353–375.

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology, 23*, 760–771.

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

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36*, 438–448.

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

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

Cook, G. I., Marsh, R. L., Hicks, J. L., & Martin, B. A. (2006). Fan effects in event-based prospective memory. *Memory, 14*, 890–900.

Coppens, L. C., Verkoeijen, P. P. J. L., & Rikers, R. M. J. P. (2011). Learning Adinkra symbols: The effect of testing. *Journal of Cognitive Psychology, 3*, 351–357.

Cyr, A.-A., & Anderson, N. D. (2015). Mistakes as stepping stones: Effects of errors on episodic memory among younger and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 841–850.

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross (Ed.). *The psychology of learning and motivation: Advances in research and theory* (Vol. 53, pp. 63–147). San Diego, CA: Elsevier.

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.

Finn, B., & Roediger, H. L. III, (2011). Enhancing retention through reconsolidation: Negative emotional arousal following retrieval enhances later recall. *Psychological Science, 22*, 781–786.

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., & Zwaan, R. A. (2014). The benefit of retrieval practice over elaborative restudy in primary school vocabulary learning. *Journal of Applied Research in Memory & Cognition, 3*, 177–182.

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition, 40*, 505–513.

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 801–812.

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory, 19*, 290–304.

Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language, 69*, 151–164.

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition, 40*, 514–527.

Jaeger, A., Eisenkraemer, R. E., & Stein, L. M. (2015). Test-enhanced learning in third-grade children. *Educational Psychology, 35*, 513–521.

Kang, S. H. K. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory & Cognition, 38*, 1009–1017.

Kang, S. H. K., & Pashler, H. (2014). Is the benefit of retrieval practice modulated by motivation? *Journal of Applied Research in Memory & Cognition, 3*, 7–12.

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language, 67*, 17–29.

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language, 62*, 227–239.

Karpicke, J. D., Blunt, J. R., Smith, M. A., & Karpicke, S. S. (2014). Retrieval-based learning: The need for guided retrieval in elementary-school children. *Journal of Applied Research in Memory & Cognition, 3*, 198–206.

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.). *The psychology of learning and motivation* (Vol. 61, pp. 237–284). San Diego, CA: Elsevier.

Kelly, J. W., Carpenter, S. K., & Sjolund, L. A. (2015). Retrieval enhances route knowledge acquisition, but only when movement errors are prevented. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 1540–1547.

Knight, J. B., Ball, B. H., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. *Journal of Memory and Language, 66*, 731–746.

Kole, J. A., & Healy, A. F. (2013). Is retrieval mediated after repeated testing? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 462–472.

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.

Kornell, N., & Vaughn, K. E. (in press). How retrieval attempts affect learning: A review and synthesis. *Psychology of Learning & Motivation*.

Kubik, V., Söderlund, H., Nilsson, L.-G., & Jönsson, F. U. (2014). Individual and combined effects of enactment and testing on memory for action phrases. *Experimental Psychology, 61*, 347–355.

Kuo, T.-M., & Hirshman, E. (1997). The role of distinctive perceptual information in memory: Studies of the testing effect. *Journal of Memory and Language, 36*, 188–201.

Lehman, M., & Karpicke, J. D. (in press). Elaborative retrieval: Do semantic mediators improve memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1787–1794.

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

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory & Cognition, 1*, 18–26.

Moeser, S. D. (1979). The role of experimental design in investigations of the fan effect. *Journal of Experimental Psychology: Human Learning & Memory, 5*, 125–134.

Myers, J. L., O'Brien, E. J., Balota, D. A., & Toyofuku, M. L. (1984). Memory search without interference: The role of integration. *Cognitive Psychology, 16*, 217–242.
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*, 402–407.
Parkin, A. J. (1980). Levels of processing and the cue overload principle. *Quarterly Journal of Experimental Psychology, 32*, 427–434.
Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*, 531–536.
Peterson, D. J., & Mulligan, N. W. (2013). The negative testing effect and multifactor account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1287–1293.
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.
Radvansky, G. A., Zacks, R. T., & Hasher, L. (1996). Fact retrieval in younger and older adults: The role of mental models. *Psychology and Aging, 11*, 258–271.
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*, 619–633.
Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255.
Roediger, H. L., III, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. P. Mestre & B. H. Ross (Eds.). *The psychology of learning and motivation: Cognition in education* (Vol. 55, pp. 1–36). San Diego, CA: Elsevier.
Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 233–239.
Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*, 1432–1463.
Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*, 76–80.
Smith, E. E., Adams, N., & Schorr, D. (1978). Fact retrieval and the paradox of interference. *Cognitive Psychology, 10*, 438–464.
Thios, S. J., & D'Agostino, P. R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning and Verbal Behavior, 15*, 529–536.
Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology, 56*, 252–257.
Vaughn, K. E., & Rawson, K. A. (2011). Diagnosing criterion level effects on memory: What aspects of memory are enhanced by repeated retrieval? *Psychological Science, 22*, 1127–1131.
Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin & Review, 19*, 1–7.
Vaughn, K. E., & Rawson, K. A. (2014). Effects of criterion level on associative memory: Evidence for associative asymmetry. *Journal of Memory and Language, 75*, 14–26.
Vaughn, K. E., Rawson, K. A., & Pyc, M. A. (2013). Repeated retrieval practice and item difficulty: Does criterion learning eliminate item difficulty effects? *Psychonomic Bulletin & Review, 20*, 1239–1245.
Watkins, O. C., & Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning & Memory, 1*, 442–452.
Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior, 16*, 465–478.
Wilson, M. (1988). The MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers, 20*, 6–10.
Yan, V. X., Yu, Y., Garcia, M. A., & Bjork, R. A. (2014). Why does guessing incorrectly enhance, rather than impair, retention? *Memory & Cognition, 42*, 1373–1383.
Zaromb, F. M., & Roediger, H. L. III, (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition, 38*, 995–1008.