Mem Cogn
DOI 10.3758/s13421-014-0477-z

Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis

Katherine A. Rawson · Kalif E. Vaughn · Shana K. Carpenter

Abstract Despite the voluminous literatures on testing effects and lag effects, surprisingly few studies have examined whether testing and lag effects interact, and no prior research has directly investigated why this might be the case. To this end, in the present research we evaluated the *elaborative retrieval hypothesis* (ERH) as a possible explanation for why testing effects depend on lag. Elaborative retrieval involves the activation of cue-related information during the long-term memory search for the target. If the target is successfully retrieved, this additional information is encoded with the cue–target pair to yield a more elaborated memory trace that enhances target access on a later memory test. The ERH states that the degree of elaborative retrieval during practice is greater when testing takes place after a long rather than a short lag (whereas elaborative retrieval during restudy is minimal at either lag). Across two experiments, final-test performance was greater following practice testing than following restudy only, and this memorial advantage was greater with long-lag than with short-lag practice. The final test also included novel cue conditions used to diagnose the degree of elaborative retrieval during practice. The overall pattern of performance in these conditions provided consistent evidence for the ERH, with more extensive elaborative retrieval during long- than during short-lag practice testing.

K. A. Rawson (✉)
Department of Psychology, Kent State University, P.O. Box 5190, Kent, OH 44242-0001, USA
e-mail: krawson1@kent.edu

K. E. Vaughn
Williams College, Williamstown, MA, USA

S. K. Carpenter
Iowa State University, Ames, IA, USA

Published online: 25 October 2014

Keywords Testing effect · Lag effect · Elaborative retrieval hypothesis · Memory

The testing effect is one of the most robust effects in memory research. More than 100 years of research has established that taking a test is more effective than restudy for improving subsequent memory (for recent reviews, see Carpenter, 2012; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Rawson & Dunlosky, 2011; Roediger, Putnam, & Smith, 2011). A sizeable body of research has also demonstrated that memory is enhanced when multiple learning trials are distributed across time rather than completed in close succession (i.e., with a longer vs. a shorter lag between presentations of a given item; Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Dunlosky et al., 2013). Within the overlapping portion of these two literatures, surprisingly few studies have examined the extent to which the benefits of testing depend on lag. Thus, it remains an open question whether testing effects are larger at long than at short lags, and if so, *why* this effect occurs. Accordingly, the goals of the present research were to investigate the extent to which lag moderates the advantage of testing over restudying, and to evaluate a theoretical explanation for why this effect occurs.

To this end, we first briefly summarize the research from the voluminous literature on testing effects that is most relevant for the present purposes, including a recent theoretical account of testing effects (the *elaborative retrieval hypothesis*, or ERH). Next, we describe the available prior research concerning whether the benefit of testing over restudy depends on lag. We then report experiments designed to make three key contributions to the extant literatures on testing and lag effects: (1) an evaluation of ERH as an explanation of testing effects, (2) new empirical evidence concerning the extent to which testing effects are moderated by lag, and most importantly, (3) an evaluation of the extent to which the ERH can explain why the benefit of testing depends on lag.

Relevant prior research on testing effects

Substantial research has shown improvement in memory when initial study is followed by a practice test versus no practice test (e.g., Butler & Roediger, 2008; Glover, 1989; Kang, McDermott, & Roediger, 2007), by more versus fewer practice tests (e.g., Pyc & Rawson, 2009; Vaughn & Rawson, 2011), and by practice testing instead of another encoding task that does not involve testing (e.g., Carpenter, Pashler, & Vul, 2006; Karpicke & Blunt, 2011; Neuschatz, Preston, Toglia, & Neuschatz, 2005; Zaromb & Roediger, 2010). Although various methods have established robust testing effects, the modal method involves comparing the advantage of testing over an equivalent amount of restudy, which is the particular effect of interest here.

Despite the wealth of testing effect research, surprisingly few theories have emerged to explain the advantage of testing over restudy. According to a recent account, the *elaborative retrieval hypothesis* (Carpenter, 2009, 2011; Carpenter & DeLosh, 2006), the presentation of a retrieval cue activates information related to that cue during search for the target in long-term memory. If the target is successfully retrieved, this additional information is encoded along with the cue–target pair, to yield a more elaborated memory trace that provides additional retrieval routes by which to access the target later. By comparison, activating elaborative information is less likely to occur when cue–target pairs are simply restudied, because no search of memory is required.

Carpenter (2009, 2011) used various methods to provide converging evidence for the ERH. We focus here on Carpenter (2011, Exp. 2), which involved the methodology adapted for the present experiments. Individuals studied weakly related word pairs (e.g., *mother–child*), followed either by a practice cued-recall test (*mother*–?) or by a restudy trial for each pair. After a 30-min filled interval, a final cued-recall test involved either the original cue word for each pair or one of two new cue words that had not been presented during initial learning. *Mediator* cues were strongly related to the original cue word but were unrelated to the target word (e.g., for *mother–child*, the mediator *father* is strongly related to *mother* but unrelated to *child*, according to association norms; Nelson, McEvoy, & Schreiber, 2004). *Related* cues were not associated with the original cue word but were weakly associated with the target (e.g., for *mother–child*, *birth* has 0 forward associative strength with *mother* and .02 forward associative strength with *child*).

On the final test, performance with the original cues demonstrated the typical advantage of testing over restudy. For evaluating the ERH, performance in the mediator- and related-cue conditions was of greater interest. Given the strong preexisting associations between the original and mediator cue words (e.g., *mother–father* has forward associative strength of .60), the mediator cues reflect elaborative information that would likely be activated by the original cue words during practice cued recall, and then encoded along with the retrieved targets. If so, the mediator words would provide effective cues for retrieving targets on the final test. In contrast, given the zero association between the original and related cue words, the related words would less likely be activated by the original cues during practice cued recall, and thus the related cues would be less effective for retrieving the targets on the final test. Note that in the restudy condition, the relative effectiveness of the mediator cues versus the related cues was predicted to be weaker than in the test condition, since restudy presumably does not invoke elaborative retrieval. As is shown in Fig. 1, their outcomes confirmed the overall pattern predicted by the ERH: The advantage of mediator over related cues was significantly greater in the testing than in the restudy group.

Does the testing effect depend on lag?

Testing is clearly potent for enhancing subsequent memory. Theoretical insight into what drives this effect can be gleaned by considering factors that moderate the advantage of testing. Of particular interest here, is the testing effect moderated by lag?[1] In contrast to other moderators that have been extensively explored (e.g., type of practice test, level of performance during practice, timing of final test, whether feedback is provided), not much is known about the extent to which lag moderates the benefit of testing over restudy. As we noted above, this gap seems surprising, given the voluminous literatures on testing and lag effects.

Thios and D'Agostino (1976) presented learners with simple sentences for initial study followed by sentence repetition or cued recall, with either 6 or 12 s between (Exp. 1) or four or 12 intervening sentences (Exp. 2) between trials. On an immediate final test involving free recall of the object phrases from each sentence, lag effects were obtained in the testing condition but not in the repetition condition. In contrast, the other available studies have not demonstrated interactive effects of testing and lag. Cull (2000, Exp. 2) had learners study vocabulary word pairs followed by three blocks of practice trials involving test–restudy, test only, or restudy only. Each block of trials was administered either consecutively (short lag) or separated by several minutes (long lag). On a final test

[1] Adopting terminology used in recent reviews (Cepeda et al., 2006; Delaney, Verkoeijen, & Spirgel, 2010), we distinguish between *spacing effects* and *lag effects*, which are often conflated. Practice trials for a given item can be presented consecutively (i.e., *massed*) or separated by intervening time or material (i.e., *spaced*); the *spacing effect* refers to enhanced performance for spaced over massed trials. When practice is spaced, the interval between trials for a given item (i.e., the *lag*) can also be varied. *Lag effects* refer to differences in performance for longer versus shorter lags, which is the effect of interest here.

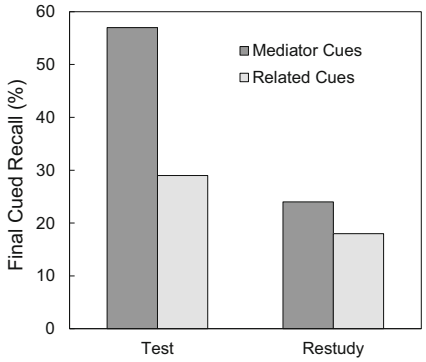

Fig. 1 From Carpenter (2011, Exp. 2), mean percentages correct on the final cued-recall test when recall of the target words was prompted with mediator cues versus related cues. Adapted from "Semantic Information Activated During Retrieval Contributes to Later Retention: Support for the Mediator Effectiveness Hypothesis of the Testing Effect," by S. K. Carpenter, 2011, *Journal of Experimental Psychology*: *Learning*, *Memory*, *and Cognition*, *37*, p. 1550. Copyright 2011 by the American Psychological Association

1 min after practice, the advantage of test–restudy over restudy only did not significantly differ for the long-lag versus the short-lag condition (a 13 % vs. 10 % advantage with *d*s of 0.52 vs. 0.41, respectively; restudy-only outperformed test-only by 30 % at both lags). One possibility is that lag effects failed to emerge because of the relatively short retention interval, on the basis of prior research showing that lag effects are often weaker (or reversed) with short than with long retention intervals (e.g., Cepeda et al., 2006; Rawson, 2012; Rawson & Kintsch, 2005).[2] Pyc and Rawson (2012b, Exp. 2) had learners study Swahili–English word pairs followed by three practice trials involving either test–restudy or restudy only, with trials separated by nine or 69 other items. Final-test performance 2 days later showed effects of both testing and lag, but no interaction. However, all learners were required to develop and report a keyword mediator after every trial for every item, which may have altered the levels of performance that would have emerged under spontaneous encoding conditions. Finally, Carpenter, Pashler, and Cepeda (2009) presented 8th-grade students with U.S. history facts for review either 1 or 16 weeks after the end of classroom instruction; the reviews involved either restudy or a cued-recall test with feedback. Performance on a final test 9 months later showed testing and lag effects but no interaction.

To our knowledge, these studies represent the only direct comparisons of testing versus restudy at different lags. Only one demonstrated that lag moderated testing effects, but aspects of the methods in others may have limited the extent to which the key interaction would emerge. Consistent with the possibility that the effects of testing are enhanced with longer versus shorter lags, other studies have demonstrated advantages of long-lag testing over short-lag testing (e.g., Kornell, 2009; Pashler, Zarow, & Triplett, 2003; Pavlik & Anderson, 2005; Pyc & Rawson, 2012a; Rawson & Dunlosky, 2013), but these studies did not include a restudy-only condition. Thus, they are silent concerning the relative advantage of testing over restudy as a function of lag, which is key for establishing that the effects reflect testing per se (rather than effects of reexposure to target information).

In sum, the extent to which the testing effect depends on lag remains an open question. Furthermore, no prior research has systematically evaluated the theoretical question of why this pattern might obtain. To the extent that lag moderates the testing effect, this outcome provides an important constraint for theory, because general theories of testing effects should be equipped to explain not only the testing effect itself, but also the factors that moderate it.

To this end, we propose a straightforward extension of the ERH as one explanation for why the benefit of testing may depend on lag: The degree of elaborative retrieval is greater when testing takes place after a long versus a short lag. Increased lag renders the target information less accessible at the time of the initial test, so the information may require a more extensive search of memory that involves a greater degree of elaboration than would be involved for a test that followed a shorter lag. This straightforward extension of the ERH predicts a greater advantage of testing over restudy following long lags than following short lags.

Overview of the present research

The goals of the present research were threefold. First, given a heightened awareness of the importance of independent replication that has recently emerged in the field (e.g., LeBel & Peters, 2011; Pashler & Harris, 2012; Roediger, 2012; Simons, 2014), one goal of the present research was to evaluate the evidence for the ERH via replication. Thus, we further evaluated the ERH as an account of testing effects by replicating and extending the basic design of Carpenter (2011). Of greater interest, the second goal was to examine the extent to which testing effects are moderated by lag. Third, and most importantly, we evaluated the extent to which the ERH can be extended to explain why the benefit of testing depends on lag.

We adapted the basic method developed by Carpenter (2011, Exp. 2), in which learners studied weakly related word pairs followed by practice trials involving either cued recall or restudy. The new variable of lag was introduced by repeating these practice trials after a long or a short lag. On a delayed final cued-recall test, recall of the target words was prompted

[2] Cull's (2000) Experiments 3 and 4 involved longer retention intervals and compared lags of minutes versus days. Unfortunately, these outcomes are not readily interpretable, due to several methodological limitations (in brief, retention interval and lag were confounded, such that the retention intervals were 6 days longer for short- than for long-lag conditions; participants completed the practice trials without supervision outside the lab; and performance was consistently at or near ceiling in one or more of the conditions).

either with the original cue words or with one of two types of new cue word (mediator or related, which we described earlier). In the original-cue condition, we expected to demonstrate (a) the standard testing effect, with greater performance following testing than following restudy, and (b) the standard lag effect, with performance being greater after long than after short lags. The empirical question of greater interest concerned the extent to which testing and lag have interactive effects.

For the theoretical questions of interest here, the key outcomes concerned performance in the mediator- and related-cue conditions. Paralleling the pattern illustrated in Fig. 1, the ERH predicted that the advantage of mediator cues over related cues would be greater following testing than following restudy. To the extent that lag moderates the standard testing effect in the original-cue condition, the theoretical question of greatest interest concerned whether the ERH can explain why the benefit of testing depends on lag. If elaborative retrieval is more likely after long than after short lags, the ERH predicts a three-way interaction—that is, the two-way interaction reflecting elaborative retrieval (i.e., a greater advantage of mediator cues over related cues for testing than for restudy) will be stronger with long-lag practice than with short-lag practice.

Experiment 1

Method

Participants and design Undergraduates participating for course credit (n = 130) were randomly assigned to one of four groups, defined by the factorial combination of the lag between trials (short vs. long) and the type of practice (test vs. restudy). Type of cue on the final test (original, mediator, or related) was manipulated within participants. Our targeted sample size was 128, on the basis of an a priori power analysis conducted using G*Power 3.1.5 (Faul, Erdfelder, Buchner, & Lang, 2009) for the primary outcome of greatest interest (the three-way interaction involving mediator vs. related cue, short vs. long lag, and testing vs. restudy), with power set at .80 to detect a small-to-medium-size effect (f = 0.15), assuming a correlation among repeated measures of .50 and α = .05.

Materials and procedure The materials included 36 English word pairs, including the 16 pairs used in Carpenter (2011, Exp. 2) and 20 additional pairs with similar characteristics (the full list is included in Appendix A). The cue–target pairs had a mean forward associative strength of .032 (Nelson et al., 2004). For each pair, we selected a *mediator* word that was strongly related to the cue word (mean associative strength from cue to mediator = .625) but that had 0 forward associative strength with the target word or with targets from any other pairs. For each pair, we also selected a *related* word that was weakly associated with the target (mean = .031) but that had 0 forward associative strength with any other target words.

All tasks and instructions were administered by computer. Participants were informed that they would learn word pairs and would later take a memory test, and the learning tasks appropriate to each group were briefly described. In all four groups, each item appeared once in each of six blocks of learning trials. In each block, items were presented one at a time in random order for 6 s each. In the *restudy* groups, all six blocks involved a study trial (SSSSSS). In the *test* groups, odd-numbered blocks involved study trials, and even-numbered blocks involved cued-recall trials (STSTST). On study trials, the cue word was presented on the left side of the screen, and the target word was underlined and presented on the right side of the screen. On test trials, the cue word was presented on the left, "???" was presented on the right, and a text box was presented below. Participants were told that they had 6 s to type in the underlined word from that pair, after which the program automatically advanced to the next item. No feedback was provided during test trials.

In the *long-lag* groups, each block of trials included all 36 items. Thus, trials for a given item were separated by 35 other items. In the *short-lag* groups, items were randomly assigned to four sets of nine pairs, and all six blocks of trials for one set were presented before proceeding to trials for the next set. Thus, the trials for a given item were separated by eight other items.

The final cued-recall test was administered 2 days later. Participants were informed that on each trial they would see a word on the left side of the screen and that they were to type in "the underlined word that comes to mind when you see this word." They were told that the word on the left might or might not be the same word that was originally presented with the underlined word, and that "whatever the word is on the left side of the screen, your job is to think of one of the underlined words that comes to mind when you see it." Test trials were presented one at a time, and responses were self-paced. On each trial, the cue word presented was either the original cue (the word paired with the target during learning) or a new cue (either the mediator word or the related word), with 12 items assigned to each cue type. The assignment of items to cue conditions was counterbalanced across participants.

As a brief but important aside, our methodological decisions concerning the amount of practice and the length of the retention interval were informed by the outcomes of a preliminary study involving less practice and a shorter retention interval. This preliminary study failed to demonstrate the usual advantages of testing over restudy and longer lag over shorter lags, and thus was not informative for the questions of interest here (i.e., whether the testing effect is moderated by lag and whether the ERH can explain the pattern). For archival

purposes, and on the basis of recent recommendations concerning the full reporting of outcomes (e.g., Cumming, 2014; Eich, 2014; Simmons, Nelson, & Simonsohn, 2011), we briefly describe this study in Appendix B but do not discuss it further.

Results and discussion

In each experiment, we report the outcomes of the inferential tests appropriate for evaluating each of our three questions of interest (for recommendations to only conduct the statistical analyses necessary to answer research questions of interest rather than omnibus analyses of variance [ANOVAs], see Judd & McClelland, 1989; Rosenthal & Rosnow, 1985; Tabachnik & Fidell, 2001). Cohen's *d* was computed using pooled standard deviations (Cortina & Nouri, 2000). We excluded from the analyses the data from four participants who did not return for Session 2 and from one participant with practice test performance more than three standard deviations below the group mean.

Mean cued recall during practice (collapsed across the three practice trials) was higher for the short-lag group (M = 95.7 %, SE = 0.7) than for the long-lag group (M = 91.7 %, SE = 1.2), $t(62) = 2.94$, $p = .005$, $d = 0.74$.

Original-cue performance: does lag moderate the testing effect? For the empirical question concerning the extent to which lag moderates the testing effect, the primary outcomes concern final-test performance in the original-cue condition (Table 1). As expected, the advantage of testing over restudy was significant, $F(1, 121) = 73.83$, $MSE = 500.90$, $p < .001$, $\eta_p^2 = .38$. The main effect of lag was also significant, $F(1, 121) = 4.60$, $p = .034$, $\eta_p^2 = .04$. Finally, inspection of the pattern in Table 1 indicates that the advantage of testing over restudy was larger in the long-lag group (a 41 % difference, $d = 1.67$) than in the short-lag group (a 28 % difference, $d = 1.39$). The interaction did not reach significance, $F(1, 121) = 2.46$, $p = .119$, $\eta_p^2 = .02$, likely reflecting the low level of achieved power to detect an effect that was smaller than anticipated (achieved power computed in G*Power 3.1.5 = .35).[3] The smaller effect may have been due in part to performance being near ceiling in the long-lag testing group. To foreshadow, Experiment 2 involved a larger sample size to increase power to detect this interaction, and indeed, this resulted in a significant interaction, in which the effects of testing over restudying were greater with long than with short lags.

Table 1 Mean percentages correct on the final cued-recall test when recall of the target words was prompted with original cues in Experiments 1–2

	Long Lag	Short Lag
Experiment 1		
Test	95.4 (1.8)	80.6 (2.7)
Restudy	54.7 (6.1)	52.4 (4.4)
Experiment 2		
Test (STSTST)	83.1 (3.2)	75.0 (2.9)
Test (SSSTTT)	83.3 (2.8)	70.0 (3.2)
Restudy	54.6 (4.6)	56.3 (4.0)

The parenthetical information included with the group labels in Experiment 2 indicates what type of trial was involved in each block (S = study, T = test). Standard errors of the means are reported in parentheses with each mean value

New-cue performance: evaluating the elaborative retrieval hypothesis Concerning replication of the key outcomes reported by Carpenter (2011; see Fig. 1), the relevant comparisons concern final-test performance in the mediator- and related-cue conditions as a function of practice group (test vs. restudy, collapsed across lags). Replicating the pattern reported by Carpenter (2011), the advantage of mediator over related cues was significantly greater in the testing group (26.7 % vs. 17.1 %, $d = 0.55$) than in the restudy group (8.1 % vs. 7.7 %, $d = 0.03$). A 2 (Practice Group: test or restudy) × 2 (Cue: mediator or related) ANOVA revealed main effects of practice group and cue [$F(1, 123) = 38.62$, $MSE = 317.92$, $p < .001$, $\eta_p^2 = .24$; $F(1, 123) = 12.31$, $MSE = 127.96$, $p = .001$, $\eta_p^2 = .09$, respectively] and a significant interaction [$F(1, 123) = 10.39$, $MSE = 127.96$, $p = .002$, $\eta_p^2 = .08$].

Of greatest interest, our third question concerned whether the ERH can be extended to explain why testing effects depend on lag. This question hinged on the assumption that we would observe a significant interaction between testing and lag in the original-cue condition. Although the anticipated interaction was not significant, the numerical pattern was consistent with the predicted outcome, and thus one might reasonably expect the patterns predicted by the ERH to emerge in the new-cue conditions.

To revisit, if elaborative retrieval is more likely to occur after long than after short lags, the ERH predicts a three-way interaction: The two-way interaction indicating elaborative retrieval (i.e., a greater advantage of mediator over related cues for testing than for restudy) will be stronger following

[3] McEldoon, Durkin, and Rittle-Johnson (2013) noted that effect sizes and observed power "should be considered when interpreting the practical significance of results, and relying too heavily on *p*-values may lead to misguided interpretations . . . limited power can be a rival explanation of statistically non-significant findings, and one must be careful not to falsely reject the alternative hypothesis" (p. 622). Also following recent recommendations (Simmons et al., 2011), we elected to terminate data collection with the intended sample size indicated by a priori power analyses and then to conduct a replication study (Exp. 2), rather than adding data to Experiment 1 in an attempt to reach statistical significance. When adopting this approach, Simmons et al. recommended that readers "should be more tolerant of imperfections in results. . . . Underpowered studies with perfect results are the ones that should invite extra scrutiny."

long-lag than following short-lag practice. First, consider the outcomes in the long-lag group, reported in the left panel of Fig. 2. A 2 (mediator or related cue) × 2 (test or restudy) ANOVA revealed main effects of cue [$F(1, 59) = 14.44$, $MSE = 168.67$, $p < .001$, $\eta_p^2 = .20$] and practice [$F(1, 59) = 30.08$, $MSE = 389.15$, $p < .001$, $\eta_p^2 = .34$], as well as a significant interaction [$F(1, 59) = 3.48$, $p = .047$, $\eta_p^2 = .07$]. As predicted, the advantage of mediator over related cues was larger in the long-lag testing group ($d = 0.76$) than in the long-lag restudy group ($d = 0.26$). In the short-lag group (right panel of Fig. 2), only the main effect of practice and the interaction were significant [cue, $F < 1$; practice, $F(1, 62) = 15.49$, $MSE = 160.74$, $p < .001$, $\eta_p^2 = .20$; interaction, $F(1, 62) = 8.30$, $MSE = 78.59$, $p = .005$, $\eta_p^2 = .12$]. The advantage of mediator over related cues was modest in the short-lag testing group ($d = 0.43$) and unexpectedly reversed in the short-lag restudy group ($d = –0.42$).[4] In part due to this potentially aberrant value, the interaction in the short-lag group was relatively strong. As a result, not surprisingly, the three-way interaction of cue, lag, and type of practice was not significant, $F < 1$.

Nonetheless, the qualitative pattern of outcomes across both panels of Fig. 2 is consistent with the pattern predicted by the ERH. As we mentioned earlier, statistical power issues could have obscured the emergence of the interaction between lag and testing and the three-way interaction implicating the role of elaborative retrieval. A key purpose of Experiment 2, therefore, was to replicate this basic design under stronger conditions involving a larger sample.

Experiment 2

Experiment 1 demonstrated the standard testing effect in the original-cue condition. We also observed the standard lag effect, with greater final-test performance in the original-cue condition with long than with short lags. Additionally, the overall pattern of outcomes was consistent with the expectation that lag would moderate the testing effect (with a larger advantage of testing over restudy with long than with short lags), but this interactive effect was somewhat smaller than anticipated. Concerning the pattern of performance in the new-cue conditions, we replicated the key outcomes of Carpenter (2011), and the numerical trends were consistent with a greater degree of elaborative retrieval with long-lag than with short-lag retrieval practice. Given these promising initial findings and the likely possibility that statistical power issues could have obscured the emergence of effects in Experiment 1, an important goal of Experiment 2 was to replicate these key outcomes with a larger sample (for recent emphasis on the importance of replicating novel findings, see LeBel & Peters, 2011; Ledgerwood & Sherman, 2012; Pashler & Harris, 2012; Roediger, 2012; Schmidt, 2009).

Experiment 2 also included a new condition designed to rule out an alternative interpretation of the outcomes that currently provide tentative support for the ERH as an account of why lag moderates the testing effect. The ERH assumes that the advantage of testing over restudy (and, by extension here, the advantage of long-lag over short-lag testing) arises from the additional elaborative processing that takes place during retrieval attempts. However, an alternative explanation is that the patterns observed in Experiment 1 reflected enhanced elaborative processing that took place during the restudy opportunities that followed the initial blocks of retrieval practice (i.e., in Blocks 3 and/or 5 of the STSTST schedule).

For example, Arnold and McDermott (2013b) presented Russian–English word pairs for initial study followed by either one or five practice cued-recall tests prior to subsequent restudy. More versus less retrieval practice enhanced encoding during subsequent restudy, as was evidenced by enhanced recall after restudy for previously unrecalled items (for converging evidence for *test-potentiated learning* from studies involving the free recall of picture names or word lists, see Arnold & McDermott, 2013a; Karpicke & Roediger, 2007). Likewise, Pyc and Rawson (2010, 2012a) demonstrated that after retrieval practice as compared to study, learners are more likely to develop effective mediators to link cue–target pairs during subsequent restudy (particularly after retrieval failures). Thus, the participants in Experiment 1 could have engaged in various types of elaborative processing during the study trials that followed test trials.

Note that the kind of elaborative processing assumed to take place by the ERH during retrieval, and the kind of elaborative processing that may underlie test-potentiated learning during subsequent restudy opportunities, could both contribute to the advantage of testing over restudy. To estimate the extent to which the key outcomes in Experiment 1 reflected test-potentiated learning during restudy rather than during retrieval, Experiment 2 also included a testing group in which all of the retrieval practice occurred after restudy (SSSTTT).

Method

Undergraduates participating for course credit ($n = 311$) were randomly assigned to one of six groups, defined by the factorial combination of lag between trials (short vs. long) and practice schedules (SSSSSS, STSTST, or SSSTTT). The type of cue used on the final test (original, mediator, or related) was manipulated within participants.

[4] In principle, on the basis of the theoretical assumptions of the ERH outlined here, one would expect lower performance for mediator cues in the short-lag restudy condition than in the other seven conditions. However, given that neither Carpenter (2011) nor the present Experiment 2 demonstrated a reversal, we assume that this outcome reflects noise.

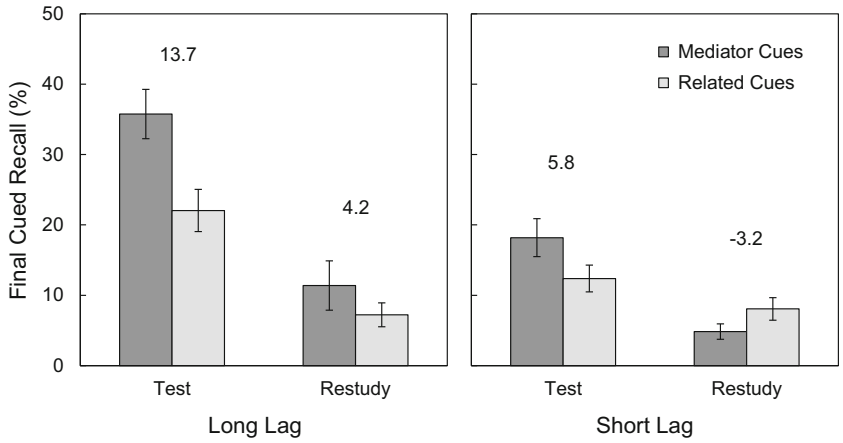

Fig. 2 Mean percentages correct on the final cued-recall test when recall of the target words was prompted with mediator cues versus related cues in Experiment 1. Error bars are standard errors of the means. Values above the bars report the numerical difference between each pair of means

Our targeted sample size was 280, on the basis of an a priori power analysis (G*Power 3.1.5; Faul et al., 2009) for the primary outcome of greatest interest (the three-way mixed-factor interaction involving mediator vs. related cue, short vs. long lag, and testing vs. restudy). For this analysis, we set power at .80 and α = .05 to detect a small effect (f = 0.10), assuming a correlation among repeated measures of .50, which indicated a sample size of 280. We then examined whether the targeted sample of 280 would be sufficient for examining the extent to which lag moderated the testing effect in the original-cue condition. For this two-way, between-subjects interaction (short vs. long lag and testing vs. restudy), a sample of 280 with power set at .80 and α = .05 had sufficient sensitivity to detect a small- to medium-sized effect (f = 0.17). The other analysis of interest concerned comparison of the STSTST and SSSTTT groups, to evaluate the extent to which the outcomes reflected elaborative processing taking place during study trials rather than during retrieval. Our expectation was that elaborative processing during subsequent restudy contributed minimally to the overall pattern of interest, and thus, functionally, we expected a minimal difference between these two testing groups. To afford sufficient power to detect even a small effect between these two groups, we planned a slight increase in the proportions of participants randomly assigned to the testing groups (50 in each of the four testing groups and 40 in each of the two restudy groups). For the three-way mixed-factor interaction involving comparison of the testing groups (mediator vs. related cue, short vs. long lag, and SSSTTT vs. STSTST), the targeted sample of 200 with power set at .80 and α = .05 had sufficient sensitivity to detect a small effect (f = 0.12). Finally, although our targeted sample size was 280, we completed a modest amount of oversampling (n = 311) to allow for some attrition.

The materials were the same as in Experiment 1. The procedures in the SSSSSS and STSTST groups were also the same. The procedure in the SSSTTT group was similar to that of the STSTST procedure, except that all three blocks of study trials were completed prior to the three blocks of cued-recall trials. All participants were asked to return 2 days later to complete the final cued-recall test.

Results and discussion

We excluded from the analyses the data for 18 participants who did not return for Session 2 and for two participants with practice test performance more than three standard deviations below the group mean.

Retrieval success during practice was relatively high in all four testing groups (STSTST: 96.7 %, SE = 0.4, for short lag and 83.1 %, SE = 1.4, for long lag; SSSTTT: 95.4 %, SE = 1.0, for short lag and 86.2 %, SE = 2.1, for long lag). The main effect of lag was significant, $F(1, 202) = 68.48$, $p < .001$, $\eta_p^2 = .25$, whereas the effect of testing schedule was not ($F < 1$).

Original-cue performance: does lag moderate the testing effect? Performance in the original-cue condition is reported in Table 1. An initial ANOVA comparing the two testing schedules indicated no main effect or interaction (Fs < 1.58). Accordingly, we collapsed across testing schedules for subsequent analyses. As expected, the advantage of testing over restudy was significant [$F(1, 287) = 56.52$, MSE = 568.72, $p < .001$, $\eta_p^2 = .17$], and the main effect of lag approached significance [$F(1, 287) = 2.97$, $p = .086$, $\eta_p^2 = .01$]. Of greater interest, the interaction between testing and lag was significant, $F(1, 287) = 5.06$, $p = .025$, $\eta_p^2 = .02$. The advantage of testing over restudy was larger in the long-lag group (a 30 % difference, d = 1.22) than in the short-lag group (a 16 % difference, d = 0.71). Thus, consistent with the pattern observed in Experiment 1, lag moderated the testing effect.

New-cue performance: evaluating the elaborative retrieval hypothesis Concerning replication of the key outcomes

reported by Carpenter (2011), the relevant comparisons concerned final-test performance in the mediator- and related-cue conditions as a function of practice group (test vs. restudy, collapsed across lags). The advantage of mediator over related cues was significantly greater in the testing group (21.0 % vs. 14.7 %, $d = 0.37$) than in the restudy group (8.9 % vs. 9.0 %, $d = –0.01$). A 2 (test or restudy) × 2 (mediator or related cue) ANOVA revealed main effects of practice group and cue [$F(1, 289) = 25.02$, $MSE = 378.30$, $p < .001$, $\eta_p^2 = .08$; $F(1, 289) = 10.10$, $MSE = 114.96$, $p = .002$, $\eta_p^2 = .03$, respectively], as well as a significant interaction [$F(1, 289) = 10.75$, $p = .001$, $\eta_p^2 = .04$].

Of greater interest, our third question concerned whether the ERH can explain why the benefit of testing depends on lag. To revisit, if elaborative retrieval is more likely after long than after short lags, the ERH predicts that the pattern indicating elaborative retrieval (i.e., a greater advantage of mediator over related cues for testing than for restudy) will be stronger in the long-lag group than in the short-lag group. The outcomes are reported in Fig. 3. In the long-lag group, a 2 (mediator or related cue) × 2 (test or restudy) ANOVA revealed main effects of cue [$F(1, 147) = 10.69$, $MSE = 135.08$, $p = .001$, $\eta_p^2 = .07$] and practice [$F(1, 147) = 19.83$, $MSE = 476.45$, $p < .001$, $\eta_p^2 = .12$], and a significant interaction [$F(1, 147) = 11.54$, $p = .001$, $\eta_p^2 = .07$]. As predicted, the advantage of mediator over related cues was larger in the long-lag testing group ($d = 0.50$) than in the long-lag restudy group ($d = –0.01$). In the short-lag group, only the main effect of practice was significant [$F(1, 140) = 8.06$, $MSE = 217.74$, $p = .005$, $\eta_p^2 = .05$; other Fs < 1.29]. The advantage of mediator over related cues was small in the short-lag testing group ($d = 0.22$), and nonexistent in the short-lag restudy group ($d = 0.00$). Confirming that the interactive pattern was stronger in the long-lag than in the short-lag group, the three-way interaction in a 2 (cue) × 2 (lag) × 2 (practice) ANOVA approached significance, $F(1, 287) = 3.48$, $MSE = 111.37$, $p = .06$, $\eta_p^2 = .01$; all other Fs < 3.12. Thus, the overall pattern of outcomes aligned with the predictions of the ERH.

Concerning the evidence for greater elaboration during long-lag than during short-lag testing, to what extent did this reflect elaboration during subsequent restudy opportunities rather than during retrieval per se? As is reported in Table 2, minimal differences between testing groups strongly suggests that the locus of elaboration is primarily during retrieval rather than during restudy. A 2 (cue) × 2 (lag) × 2 (testing group) ANOVA revealed no significant effects or interactions involving testing group [Fs < 2.42, ps > .122, $\eta_p^2 \leq .01$]. The only significant interaction was between cue and lag [$F(1, 202) = 10.03$, $MSE = 123.37$, $p = .002$, $\eta_p^2 = .05$]; the advantage of mediator over related cues was significantly greater with long-lag testing ($d = 0.50$) than with short-lag testing ($d = 0.22$). This outcome provides additional evidence that elaboration during retrieval occurs to a greater extent with longer than with shorter lags.

The minimal evidence for test-potentiated learning seems at first inconsistent with findings in the prior research (e.g., Arnold & McDermott, 2013b; Pyc & Rawson, 2012b). However, the extent to which test-potentiated learning occurs may depend on the level of performance during practice, which was much higher here than in the previous studies. The effects reported by Arnold and McDermott (2013b) and Pyc and Rawson (2012b) were particularly pronounced following retrieval failures, and other research has demonstrated larger benefits of restudy for previously incorrect items than for previously correct items (e.g., Pashler, Cepeda, Wixted, & Rohrer, 2005; Vojdanoska, Cranney, & Newell, 2010).

General discussion

Despite the voluminous literatures on testing effects and lag effects, surprisingly few studies have examined the degree to

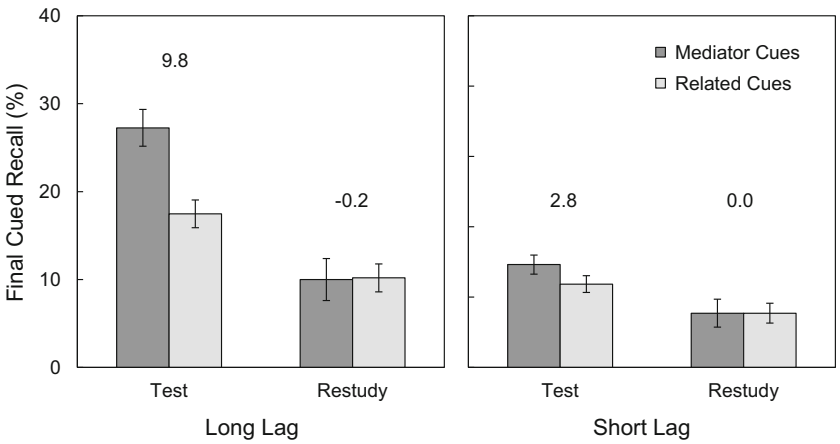

Fig. 3 Mean percentages correct on the final cued-recall test when recall of the target words was prompted with mediator cues versus related cues in Experiment 2. Error bars are standard errors of the means. Values above the bars report the numerical difference between each pair of means

Table 2 Mean percentages correct on the final cued-recall test when the recall of target words was prompted with new cues, for the two testing groups in Experiment 2

	Mediator Cues	Related Cues
Testing With Subsequent Restudy (STSTST)		
Long lag	27.7 (2.8)	14.9 (1.9)
Short lag	16.7 (1.9)	13.5 (1.9)
Testing Without Subsequent Restudy (SSSTTT)		
Long lag	26.8 (3.1)	20.1 (2.5)
Short lag	12.7 (1.9)	10.3 (1.5)

The parenthetical information included with each group label indicates what type of trial was involved in each block (S = study, T = test). Standard errors of the means are reported in parentheses with each mean value

which these two factors interact, and no prior research has directly investigated why this might be the case. In both experiments reported here, the memorial advantage of testing over restudy was greater following practice that involved longer rather than shorter lags between trials. Both experiments also replicated the key outcomes of Carpenter (2011) to further establish the involvement of elaborative retrieval in testing effects. Importantly, the pattern of outcomes in the new-cue conditions indicated that the ERH can be extended to explain why the benefit of testing depends on lag. According to the ERH, presentation of a retrieval cue initiates a long-term memory search that involves activation of related information en route to the target. This information is encoded along with the successfully retrieved target, which aids future memory by providing additional cues that can facilitate retrieval of the target (see, e.g., Carpenter, 2009, 2011). As it was originally formulated, the ERH stated that elaborative retrieval is more likely to occur during testing than during restudy. Here, we further proposed that conditions that render the target less accessible at the time of initial retrieval, such as increased lag, may require a more extensive search of memory, and thus may enhance this elaborative process. The present research provides evidence that elaborative retrieval plays a role in the moderating effect of lag on the benefits of testing.

Although the primary outcomes of interest concerned final recall performance, secondary outcome was to provide additional evidence for the assumptions that underlie the predictions of the ERH. First, consistent with the assumption that cue-related information is activated during retrieval practice, 25 % (*SE* = 1.8) of incorrect responses during practice involved a commission error in which the mediator word was produced instead of the target. Second, the ERH's claim that elaborative retrieval is less extensive during short-lag than during long-lag testing rests on the assumption that accessing target information in long-term memory is more rapid in the former than in the latter conditions. Consistent with this assumption, first-keypress latencies for the first time that an item was correctly recalled during practice (i.e., the time in milliseconds between cue onset and the first keypress of the entry of a correct response) were faster for short-lag than for long-lag items [Exp. 1, 1,420 vs. 1,692 ms, $t(64) = 4.14$, $d = 1.04$; Exp. 2, 1,413 vs. 1,716 ms, $t(204) = 8.57$, $d = 1.19$]. A faster or less extensive search of memory affords less time or opportunity for other cue-related information to be activated and encoded along with the retrieved target. Finally, Table 3 reports the mean first-keypress latencies for correct responses in the new-cue conditions on the final test. These outcomes should be interpreted with some caution, however, given that some means are based on a small number of items (due to low levels of correct performance) and/or on a reduced number of participants (participants with 0 % correct recall in one or more conditions did not contribute a value to these cells; these missing values also preclude meaningful omnibus inferential tests). Nonetheless, the qualitative patterns largely parallel those observed in final recall performance.

The stronger lag effects for tested than for restudied items in the original-cue conditions aligns with the overall pattern that emerges from the larger literature on lag effects. Although some prior studies have demonstrated lag effects with restudy (e.g., Bjork & Allen, 1970; Elmes, Dye, & Herdelin, 1983; Gartman & Johnson, 1972; Kahana & Howard, 2005; Melton, 1970; Siegel & Kahana, 2014; Thios, 1972; Toppino & Bloom, 2002), these effects often involved free recall of word lists rather than cued recall of word pairs, were relatively modest, and were not always replicated in subsequent research (e.g., Toppino & Gracen, 1985, reported nine experiments failing to show lag effects previously reported by Glenberg,

Table 3 Mean first-keypress latencies (in milliseconds) for correct responses in the new-cue conditions on the final test, collapsed across Experiments 1 and 2

	Mediator Cues		Related Cues	
	Long Lag	Short Lag	Long Lag	Short Lag
Test	3,352 (131)	4,012 (255)	4,670 (272)	4,175 (297)
Restudy	4,237 (519)	4,196 (541)	5,426 (493)	5,127 (483)

Standard errors of the means are reported in parentheses. Trials with first-keypress latencies longer than 20 s were excluded from calculation of an individual's mean value for a given condition (on average, less than one trial per participant was excluded). Individual mean values more than three standard deviations above the group mean were excluded from the group analyses (1.3 % of all values). Outcomes should be interpreted with some caution, given that some means are based on a small number of items (due to low levels of correct performance) and/or a reduced number of participants (participants with 0 % correct recall in one or more conditions did not contribute a value to these cells; these missing values also preclude meaningful omnibus inferential tests). Additionally, latencies for the mediator- and related-cue conditions are not directly comparable, given differences in the item characteristics (e.g., word length, frequency) that may also influence response latencies

1977). A surprising number of studies have reported no effect of lag with restudy (e.g., Ausubel, 1966; Boywitt & Brandt, 2012; Braun & Rubin, 1998; Challis, 1993; Cuddy & Jacoby, 1982; Dellarosa & Bourne, 1985; Elmes, Greener, & Wilkinson, 1972; Rose, 1984; Underwood, Kapelak, & Malmi, 1976; Waugh, 1970; Wenger, 1979). In contrast, lag effects for testing conditions are quite robust (e.g., Bahrick, 1979; Bahrick & Hall, 2005; Bell, Kawadri, Simone, & Wiseheart, 2014; Bloom & Shuell, 1981; Fishman, Keller, & Atkinson, 1968; Glover, 1989; Jacoby, 1978; Kornell, 2009; Pashler et al., 2003; Pavlik & Anderson, 2005; Peterson & Gentile, 1965; Pyc & Rawson, 2007, 2009; Rawson & Dunlosky, 2013; Rohrer, 2009; Simone, Bell, & Cepeda, 2013; Sobel, Cepeda, & Kapler, 2011).

Alternative interpretations

According to the ERH, the performance advantage for long-lag testing reflects increased activation and encoding of the mediator along with the target during retrieval practice, thus providing an additional retrieval route at final test (mediator ➔ target). An alternative interpretation of this outcome is that presentation of the mediator cue may instead activate the cue word, which participants then use to activate the target word (mediator ➔ cue ➔ target). If so, testing effects in the mediator-cue condition would not reflect stronger mediator–target connections, but rather stronger cue–target associations for tested than for restudied items. This account is plausible, given the average backward associative strength of .18 between the cue–mediator pairs used by Carpenter (2011) and in the present research. However, Brennan, Cho, and Neely (2013) recently reported outcomes weighing against this possibility. They replicated Carpenter's (2011) study using her materials and extended the findings by also developing a new set of materials with a backward associative strength of 0 for cue–mediator pairs. Both sets of items showed similar testing effects in the original-cue condition and, importantly, similar advantages of testing over restudy in the mediator-cue condition. By extension, these outcomes suggest that the moderating effects of lag demonstrated here are unlikely to be accounted for by traversal of backward associations. Nonetheless, further research using methods other than the new-cue methodology (e.g., the false recognition method; Carpenter, 2011, Exp. 1) would be valuable for providing converging evidence for the role of elaborative retrieval in the moderating effects of lag on test-enhanced learning.

To manipulate lag while keeping the number of items to be learned constant, we adopted the modal method used in studies that explore the effects of lag—namely, dividing items into smaller subsets that are practiced in blocks. An inherent feature of this method is that retrieval practice for one subset of items occurs before the next subset is presented and practiced, and recent research has shown that such *interim testing* can enhance the learning and retention of subsequently presented information (e.g., Szpunar, McDermott, & Roediger, 2008; Wissman, Rawson, & Pyc, 2011). These outcomes suggest that the present outcomes may have underestimated the benefit of long- versus short-lag testing if our interim tests enhanced the learning of short-lag items. To the extent that interim tests delineate different list contexts, other research suggests that reinstating the context of previously learned information can impair memory for more recently learned information (e.g., Sahakyan & Hendricks, 2012). If so, the interim testing inherent to instantiating short-lag practice may have had offsetting positive and negative effects. However, both of these effects have been examined under conditions in which list context was particularly important for performance (i.e., free recall of multiple word lists), whereas list context likely played less of a role in the present conditions (i.e., cued recall of paired associates).

Theoretical implications

Although the present research was specifically designed to evaluate the ERH, the finding that lag moderates the testing effect also has important implications for other theories of testing effects. To the extent that general theories of testing effects should be equipped to explain not only why testing benefits memory but also the factors that moderate these effects, it is useful to consider the present findings in light of other explanations that have been offered to explain why testing enhances memory. For example, according to the *bifurcated distribution model* (BDM; Kornell, Bjork, & Garcia, 2011), items that are successfully retrieved on a practice test receive a large increment in memory strength, whereas those that are not retrieved receive no increment in memory strength. The distribution of items is thus bifurcated into two sets of items—those that are high in memory strength, and those that are low in memory strength. In contrast, all restudied items are assumed to receive an intermediate increment in memory strength. The extent to which final-test performance reveals advantages of testing over restudying depends upon (a) the degree of retrieval success during practice (i.e., the number of items receiving a large increment in memory strength) and (b) the threshold for recall on the final test (if the threshold is low—e.g., after a short retention interval—most restudied items will have sufficient strength to be recalled even if their memory strength is lower than that of the tested items). The BDM could accommodate the finding that lag moderates testing effects by also assuming that the increment in memory strength for successfully retrieved items is greater after long than after short lags. Although this additional assumption would account for the interactive effects of lag and testing observed in the original-cue condition, the BDM does not specify the exact nature of memory strength and thus is not well equipped to explain how a unitary

memory strength would yield the differential patterns observed in the mediator- versus the related-cue conditions.

The *episodic-context account* (ECA; Karpicke, Lehman, & Aue, 2014) assumes that items are initially encoded along with information about the temporal context in which the item was encountered. The ECA further assumes that "when the context during retrieval has changed significantly from the context during study, subjects attempt to *reinstate* the temporal context associated with the study period" (Karpicke et al., 2014, p. 258, italics in original). If study phase retrieval is successful, an item's context representation is updated to include features of both the original and the current temporal context, and thus 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 ECA's assumption that reinstatement of prior context occurs when the context during retrieval has changed significantly is consistent with the finding that lag moderates testing effects in the original-cue condition, given that the change in temporal context from study to retrieval is greater with longer than with shorter lags. In contrast, it is less obvious how the ECA would explain the patterns observed in the new-cue conditions. Given that the mediator- and related-cue words had not been presented previously, neither cue would have temporal context information associated with it to selectively benefit tested versus restudied items or to yield a greater benefit for mediator than for related cues.

With that said, the BDM, ECA, and ERH accounts are not mutually exclusive, and just as the effects of testing on learning are multifaceted (see, e.g., Roediger et al., 2011, for a review of ten different benefits of testing for enhancing learning), so too would be the explanations of these effects. Any given testing effect almost certainly involves more than one underlying mechanism, and any given mechanism is unlikely to play a role in all observations of testing effects. Thus, the ERH is unlikely to entirely account for testing effects or for the moderating effects of lag, since these effects very likely involve other mechanisms, as well. Furthermore, we are not claiming that elaborative retrieval is involved in all types or conditions of practice testing. For example, elaborative retrieval would seem less likely to play a role for materials in which cues do not easily afford the encoding or activation of elaborative cue-based information (e.g., Adrinka symbols or Chinese characters; Coppens, Verkoeijen, & Rikers, 2011; Kang, 2010). Elaborative retrieval may also make larger contributions during earlier stages of learning than after extended practice or overlearning (e.g., Kole & Healy, 2013). This possibility follows from the same logic described above, that elaborative retrieval is less likely when access to target information in long-term memory is relatively rapid. With extended practice, response latencies also decrease as the number of correct recalls for a given item increase (e.g., Pyc & Rawson, 2009), suggesting that the contribution of elaborative retrieval may diminish with overlearning. Elaborative retrieval would also be expected to play a minimal role when retrieval success during practice is low, given that the memorial benefit arises from encoding elaborative information along with successfully retrieved targets. Finally, the ERH was developed primarily to explain the nature of the associations between cues and targets and, thus, as it is currently formulated, would be less relevant to learning of materials that do not afford this type of associative processing.

We have focused here on one kind of elaborative processing—namely, the elaborative encoding that arises from the activation of cue-related information during a search of long-term memory for target information. Testing may also invoke other kinds of elaborative processing, above and beyond elaborative retrieval as conceptualized here. For example, closely related work has provided evidence that retrieval practice can enhance the effectiveness of mediators that are intentionally encoded during subsequent restudy opportunities (Pyc & Rawson, 2010, 2012b).

More generally, *elaborative processing* can be conceptualized as an umbrella term that encompasses conditions in which information not explicitly presented or contained within a nominal memorandum is encoded along with the memorandum. However, the quality or quantity of the elaborative information encoded may differ as a function of the encoding task, and thus, different types of study activities may have differential effects on learning. Relevant to this question, recent studies have shown that retrieval produces benefits to later memory that exceed those produced by elaborative-encoding activities such as the keyword mnemonic method during study of foreign language vocabulary (Karpicke & Smith, 2012), the generation of additional words that come to mind when studying lists of to-be-remembered words (Lehman, Smith, & Karpicke, 2014), and the linking of newly learned vocabulary words to other words with similar meanings (Goossens, Camp, Verkoeijen, Tabbers, & Zwaan, 2014). These findings suggest that, as compared to retrieval, these tasks may differ in their quality or quantity of elaborative encoding. We encourage future research that might shed additional light on the nature of elaborative types of processing that occur during encoding and retrieval, and how this processing might affect subsequent memory.

Conclusions

The present work provides the first empirical demonstration that the advantage of testing over restudy depends on lag. Furthermore, the respective literatures on testing and lag effects are sizeable, but neither effect is sufficiently well understood, with theory lagging behind empirical demonstrations of these effects. These results represent an important theoretical contribution to both of these literatures by further establishing that elaborative retrieval contributes to

testing effects, and by providing the first evidence for the involvement of elaborative retrieval in the moderation of those effects as a function of lag.

Author note The research reported here was supported by a James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind and Behavior Collaborative Award.

Appendix A: Items used in Experiments 1 and 2

Original Cues	Target Words	Mediator Cues	Related Cues	Associative Strength		
				C-M	C-T	R-T
mother	child	father	birth	.597	.010	.015
prescription	doctor	drug	hospital	.477	.034	.027
soil	earth	dirt	continent	.717	.040	.041
dusk	evening	dawn	morning	.609	.042	.047
donor	heart	blood	liver	.524	.042	.041
weapon	knife	gun	axe	.592	.075	.046
sonnet	music	poem	dancer	.471	.059	.052
employment	office	job	government	.605	.020	.024
trash	paper	garbage	ink	.526	.013	.013
vocabulary	school	words	text	.507	.013	.013
jacket	shirt	coat	hanger	.564	.013	.014
pedestrian	street	walk	neighborhood	.597	.032	.034
breeze	summer	wind	mosquito	.606	.012	.014
coffee	table	tea	banquet	.442	.020	.020
frame	window	picture	shingle	.811	.014	.014
agony	ecstasy	pain	pleasure	.649	.019	.014
alive	breathe	dead	vapor	.554	.020	.040
anatomy	biology	body	lab	.607	.028	.022
bait	hook	fish	pirate	.629	.053	.047
clock	hands	time	pray	.652	.036	.033
pepper	sneeze	salt	dust	.695	.041	.054
umbrella	dry	rain	rinse	.701	.042	.035
wallet	leather	money	boots	.630	.041	.042
vine	ivy	grape	league	.605	.020	.014
flipper	scuba	dolphin	tank	.797	.020	.016
glacier	mountain	ice	volcano	.723	.020	.022
hammer	pound	nail	knock	.800	.028	.027
rake	hoe	leaves	shovel	.622	.047	.042
king	crown	queen	jewel	.772	.016	.020
peel	potato	orange	tomato	.571	.065	.062
pork	beef	pig	cattle	.594	.042	.041
lamp	post	light	mailbox	.769	.026	.013
lather	shave	soap	whiskers	.673	.048	.073
mouse	trap	cat	cage	.543	.029	.028
tusk	tooth	elephant	pick	.660	.028	.034
antler	moose	deer	noose	.615	.027	.021

The first 16 items are the same as in Carpenter (2011, Exp. 2). C-T = cue-to-target strength. C-M = cue-to-mediator strength. R-T = related-to-target strength. The mediator-to-target strength was 0 for all pairs. The associative strength between mediator cues and all other target words was 0, and the associative strength between related cues and all other target words was also 0.

Appendix B: Preliminary study that informed the methodology used in Experiments 1 and 2

Method

Undergraduates (n = 128) were randomly assigned to one of four groups defined by lag (short vs. long) and type of practice (test vs. restudy). Final-test cue (original, mediator, or related) was manipulated within participants. The targeted sample size of 128 was based on the same power analysis reported for Experiment 1. The materials and procedure were the same as in Experiment 1, except that (1) items were presented for three blocks of learning trials (either SSS in the restudy group or STT in the test group), and (2) the final test was administered 20 min after the final block.

Results and discussion

During practice, mean cued recall was significantly greater with a short lag (85.2 %, SE = 2.8) than with a long lag (70.2 %, SE = 3.5), $t(63) = 3.32$, $p = .001$, $d = 0.82$. Concerning final-test performance in the original-cue condition (Table 4), we did not find significant effects of either testing ($F < 1$) or lag, $F(1, 124) = 2.80$, $MSE = 600.08$, $p = .097$, $\eta_p^2 = .02$ (interaction $F < 1$). In hindsight, these outcomes are consistent with prior research showing weaker or reversed lag effects with shorter retention intervals and/or low levels of practice (e.g., Cepeda et al., 2006; Pavlik & Anderson, 2005; Rawson, 2012) and weaker or reversed testing effects at short retention intervals (e.g., Congleton & Rajaram, 2012; Coppens et al., 2011; Roediger & Karpicke, 2006; Toppino & Cohen, 2009).

Table 4 Mean percentages correct on the final cued-recall test

	Long Lag		Short Lag	
	Test	Restudy	Test	Restudy
Original cues	69.7 (4.0)	72.1 (5.2)	78.7 (4.4)	77.7 (3.6)
Mediator cues	27.8 (4.4)	15.1 (3.4)	28.4 (4.6)	12.4 (3.3)
Related cues	17.7 (3.0)	10.7 (2.7)	15.1 (3.2)	9.4 (2.1)

Standard errors of the means are reported in parentheses with each mean value

With that said, a testing effect may still have been expected, given that the present experiment was a close replication of Carpenter (2011). One difference between these two studies involved set size (16 items in Carpenter, 2011, vs. 36 items here, to include enough items in each cell of the expanded design to accommodate the additional manipulation of lag). Consistent with list length effects (e.g., Cary & Reder, 2003; Ward & Tan, 2004), practice recall was lower in Experiment 1 than in Carpenter's (2011) study (78 % overall in Exp. 1 vs. 91 % in Carpenter, 2011), and testing effects depend in part on the level of retrieval success during practice (e.g., Halamish & Bjork, 2011; Kornell et al., 2011). These outcomes motivated methodological changes in Experiments 1–2 to increase successful retrieval during practice (by increasing the amount of practice) and to use more sensitive conditions for detecting testing and lag effects (by using a longer retention interval).

Final-test performance in the mediator- and related-cue conditions (Table 4) replicated the key outcomes reported by Carpenter (2011): The advantage of mediator cues over related cues was greater in the testing group (28.1 vs. 16.4 %, d = 0.50) than in the restudy group (13.8 vs. 10.1 %, d = 0.22). A 2 (Practice Group: test or restudy) × 2 (Cue: mediator or related) ANOVA revealed main effects of practice group and cue [$F(1, 126) = 11.28$, $MSE = 606.34$, $p = .001$, $\eta_p^2 = .08$; $F(1, 126) = 26.42$, $MSE = 143.07$, $p < .001$, $\eta_p^2 = .17$, respectively] and a significant interaction [$F(1, 126) = 7.09$, $MSE = 143.07$, $p = .009$, $\eta_p^2 = .05$]. Given the absence of lag effects in the original-cue condition, not surprisingly, no effect or interaction involving lag was significant in the new-cue conditions, Fs < 1.

References

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. doi:10.1037/a0029199

Ausubel, D. P. (1966). Early versus delayed review in meaningful learning. *Psychology in the Schools, 3,* 195–198.

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108,* 296–308.

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language, 52,* 566–577. doi:10.1016/j.jml.2005.01.012

Bell, M. C., Kawadri, N., Simone, P. M., & Wiseheart, M. (2014). Long-term memory, sleep, and the spacing effect. *Memory, 22,* 276–283.

Bjork, R. A., & Allen, T. W. (1970). The spacing effect: Consolidation or differential encoding? *Journal of Verbal Learning and Verbal Behavior, 9,* 567–572.

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research, 74,* 245–248.

Boywitt, C. D., & Brandt, M. (2012). The primacy effect in memory for repetitions: Evidence for the role of lag between repetitions in source monitoring. *Journal of Cognitive Psychology, 24,* 295–305.

Braun, K., & Rubin, D. C. (1998). The spacing effect depends on an encoding deficit, retrieval, and time in working memory: Evidence from once-presented words. *Memory, 6,* 37–65.

Brennan, M. K., Cho, K. W., & Neely, J. H. (2013). *The role of mediators in the testing effect in paired-associate learning.* Paper presented at the 54th Annual Meeting of the Psychonomic Society, Toronto, ON.

Butler, A. C., & Roediger, H. L., III. (2008). Feedback enhances the positive effects and reduces the negative effects multiple-choice testing. *Memory & Cognition, 36,* 604–616. doi:10.3758/MC.36.3.604

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. doi:10.1037/a0017021

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. doi:10.1037/a0024140

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

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34,* 268–276. doi:10.3758/BF03193405

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. doi:10.1002/acp.1507

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. doi:10.3758/BF03194004

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language, 49,* 231–248.

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. doi:10.1037/0033-2909.132.3.354

Challis, B. H. (1993). Spacing effects on cued-memory tests depend on level of processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 389–396. doi:10.1037/0278-7393.19.2.389

Congleton, A., & Rajaram, S. (2012). The origin of the interaction between learning method and delay in the testing effect: The roles of processing and conceptual retrieval organization. *Memory & Cognition, 40,* 528–539.

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, 23,* 351–357.

Cortina, J. M., & Nouri, H. (2000). *Effect size for ANOVA designs*. Thousand Oaks, CA: Sage.

Cuddy, L. J., & Jacoby, L. L. (1982). When forgetting helps memory: An analysis of repetition effects. *Journal of Verbal Learning and Verbal Behavior, 21,* 451–467. doi:10.1016/S0022-5371(82)90727-7

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14,* 215–235.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25,* 7–29. doi:10.1177/0956797613504966

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.

Dellarosa, D., & Bourne, L. E., Jr. (1985). Surface form and the spacing effect. *Memory & Cognition, 13,* 529–537.

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.

Eich, E. (2014). Business not as usual [Editorial]. *Psychological Science, 25,* 3–6. doi:10.1177/0956797613512465

Elmes, D. G., Dye, C. J., & Herdelin, N. J. (1983). What is the role of affect in the spacing effect? *Memory & Cognition, 11,* 144–151.

Elmes, D. G., Greener, W. I., & Wilkinson, W. C. (1972). Free recall of items presented after massed- and distributed-practice items. *American Journal of Psychology, 85,* 237–240.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41,* 1149–1160. doi:10.3758/BRM.41.4.1149

Fishman, E. J., Keller, L., & Atkinson, R. C. (1968). Massed versus distributed practice in computerized spelling drills. *Journal of Educational Psychology, 59,* 290–296.

Gartman, L. M., & Johnson, N. F. (1972). Massed versus distributed repetitions of homographs: A test of the differential-encoding hypothesis. *Journal of Verbal Learning and Verbal Behavior, 11,* 801–808.

Glenberg, A. M. (1977). Influences of retrieval processes on the spacing effect in free recall. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 282–294. doi:10.1037/0278-7393.3.3.282

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

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 and Cognition, 3,* 177–182. doi:10.1016/j.jarmac.2014.05.003

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.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17,* 649–667. doi:10.1016/S0022-5371(78)90393-6

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model comparison approach*. New York, NY: Harcourt Brace Jovanovich.

Kahana, M. J., & Howard, M. W. (2005). Spacing and lag effects in free recall of pure lists. *Psychonomic Bulletin & Review, 12,* 159–164. doi:10.3758/BF03196362

Kang, S. H. K. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory & Cognition, 38,* 1009–1017. doi:10.3758/MC.38.8.1009

Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19,* 528–558.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331,* 772–775. doi:10.1126/science.1199327

Karpicke, J. D., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57,* 151–162. doi:10.1016/j.jml.2006.09.004

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language, 67,* 17–29. doi:10.1016/j.jml.2012.02.004

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 Academic Press.

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. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23,* 1297–1317. doi:10.1002/acp.1537

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.

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology, 15,* 371–379.

Ledgerwood, A., & Sherman, J. W. (2012). Short, sweet, and problematic? The rise of the short report in psychological science. *Perspectives on Psychological Science, 7,* 60–66.

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*. doi:10.1037/xlm0000012

McEldoon, K. L., Durkin, K. L., & Rittle-Johnson, B. (2013). Is self-explanation worth the time? A comparison to additional practice. *British Journal of Educational Psychology, 83,* 615–632.

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior, 9,* 596–606.

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. doi:10.3758/BF03195588

Neuschatz, J. S., Preston, E. L., Toglia, M. P., & Neuschatz, J. S. (2005). Comparison of the efficacy of two name-learning techniques: Expanding rehearsal and name–face imagery. *American Journal of Psychology, 118,* 79–102.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 3–8. doi:10.1037/0278-7393.31.1.3

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7,* 531–536.

Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1051–1057. doi:10.1037/0278-7393.29.6.1051

Pavlik, P. I., Jr., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science, 29,* 559–586.

Peterson, L. R., & Gentile, A. (1965). Proactive interference as a function of time between tests. *Journal of Experimental Psychology, 70,* 473–478.

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. doi:10.1016/j.jml.2009.01.004

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330,* 335. doi:10.1126/science.1191465

Pyc, M. A., & Rawson, K. A. (2012a). Are judgments of learning made after correct responses during retrieval practice sensitive to lag and criterion level effects? *Memory & Cognition, 40,* 976–988.

Pyc, M. A., & Rawson, K. A. (2012b). Why is test–restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis.

Journal of Experimental Psychology: Learning, Memory, and Cognition, 38, 737–746. doi:10.1037/a0026166

Rawson, K. A. (2012). Why do rereading lag effects depend on test delay? *Journal of Memory and Language, 66,* 870–884.

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General, 140,* 283–302. doi:10.1037/a0023956

Rawson, K. A., & Dunlosky, J. (2013). Relearning attenuates the benefits and costs of spacing. *Journal of Experimental Psychology: General, 142,* 1113–1129.

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend upon time of test. *Journal of Educational Psychology, 97,* 70–80.

Roediger, H. L. III. (2012, February). Psychology's woes and a partial cure: The value of replication. *APS Observer*, *25*(2), 9, 27–29. Retrieved from www.psychologicalscience.org/index.php/publications/observer/2012/february-11-2012-observer-publications/psychology's-woes-and-a-partial-cure-the-value-of-replication.html

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249–255. doi:10.1111/j.1467-9280.2006.01693.x

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 Academic Press.

Rohrer, D. (2009). Avoidance of overlearning characterizes the spacing effect. *European Journal of Cognitive Psychology, 21,* 1001–1012.

Rose, R. J. (1984). Processing time for repetitions and the spacing effect. *Canadian Journal of Psychology, 83,* 537–550.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. Cambridge, UK: Cambridge University Press.

Sahakyan, L., & Hendricks, H. E. (2012). Context change and retrieval difficulty in the list-before-last paradigm. *Memory & Cognition, 40,* 844–860.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social science. *Review of General Psychology, 13,* 90–100.

Siegel, L. L., & Kahana, M. J. (2014). A retrieved context account of spacing and repetition effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 755–764.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. doi:10.1177/0956797611417632

Simone, P. M., Bell, M. C., & Cepeda, N. J. (2013). Diminished but not forgotten: Effects of aging on magnitude of spacing effect benefits. *Journals of Gerontology, 68B,* 674–680. doi:10.1093/geronb/gbs096

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9,* 76–80. doi:10.1177/1745691613514755

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology, 25,* 763–767.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L., III. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1392–1399. doi:10.1037/a0013082

Tabachnik, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston, MA: Allyn & Bacon.

Thios, S. J. (1972). Memory for words in repeated sentences. *Journal of Verbal Learning and Verbal Behavior, 11,* 789–793.

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., & Bloom, L. C. (2002). The spacing effect, free recall, and two-process theory: A closer look. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 437–444. doi:10.1037/0278-7393.28.3.437

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval. *Experimental Psychology, 56,* 252–257.

Toppino, T. C., & Gracen, T. F. (1985). The lag effect and differential organization theory: Nine failures to replicate. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 185–191. doi:10.1037/0278-7393.11.1.185

Underwood, B. J., Kapelak, S. M., & Malmi, R. A. (1976). The spacing effect: Additions to the theoretical and empirical puzzle. *Memory & Cognition, 4,* 391–400. doi:10.3758/BF03213195

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. doi:10.1177/0956797611417724

Vojdanoska, M., Cranney, J., & Newell, B. R. (2010). The testing effect: The role of feedback and collaboration in a tertiary classroom setting. *Applied Cognitive Psychology, 24,* 1183–1195.

Ward, G., & Tan, L. (2004). The effect of the length of to-be-remembered lists and intervening lists on free recall: A reexamination using overt rehearsal. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 1196–1210. doi:10.1037/0278-7393.30.6.1196

Waugh, N. C. (1970). On the effective duration of a repeated word. *Journal of Verbal Learning and Verbal Behavior, 9,* 587–595.

Wenger, S. K. (1979). The within-list distributed practice effect: More evidence for the inattention hypothesis. *American Journal of Psychology, 92,* 105–113.

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review, 18,* 1140–1147. doi:10.3758/s13423-011-0140-7

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. doi:10.3758/MC.38.8.995