Learning to recall: Examining recall latencies to test an intra-item learning theory of testing effects

William J. Hopper and David E. Huber

University of Massachusetts, Amherst

Author Note

Correspondence concerning this article should be addressed to William Hopper, Department of Psychological and Brain Sciences, 441 Tobin Hall, University of Massachusetts, 135 Hicks Way, Amherst, MA 01003 USA.

E-mail: whopper@psych.umass.edu

This work was supported by the National Science Foundation grant number BCS-1431147

Conflicts of interest: none

Abstract

We propose a new theory for the benefits of recall practice based on intra-item learning. On this account, retrieval cues produce an initial memory state (termed 'primary retrieval'). However, this state is incomplete and insufficient for overt recall of the item. A subsequent process, termed 'convergent retrieval', fills in any missing information through intra-item associations, allowing recall of the item. Because this occurs in a staged manner, directional learning occurs from the initially retrieved features to the subsequently retrieved features; in contrast, restudy produces less intra-item learning because restudy provides all features simultaneously. This account of the testing effect makes unique predictions regarding recall latencies. We confirmed these predictions in two experiments, examining recall latencies in free recall and cued recall. Specifically, for a final test taken immediately after a practice test that did not include accuracy feedback, restudy produced higher accuracy than test practice, but, at the same time, test practice produced faster recall than restudy. In other words, a comparison between accuracy and recall latencies suggests a process dissociation for the benefits of each type of practice. Alternative accounts of these effects were ruled out: 1) response order analyses of the free recall experiment ruled out cue-target associations; and 2) a cue-switching manipulation in the cued recall experiment (recall practice with cue A, final recall with cue B) ruled out context-target associations. According to the proposed theory, intra-item learning is narrow in one sense (i.e., unique to the cues used during practice), but robust in another sense (i.e., learning how to recall the item).

Keywords: free recall; cued recall; testing effect; recall latency

Learning to recall: Examining recall latencies to test an intra-item learning theory of testing effects

The testing effect describes the benefits of taking a practice test, including the ubiquitous finding of better long-term retention of information after a practice test as compared to restudying the same information (See Roediger & Butler, 2011 for a review). Thus, if an exam will occur in five minutes, the best strategy may be to skim-read the material, but when studying well in advance of an exam, a better strategy is to use a set of flash cards to practice recalling the material. Despite many empirical investigations and several proposed theoretical accounts, there is no universally accepted explanation for testing effects, or retrieval-based learning more broadly (Rowland, 2014). In this study we do not present a complete theory of retrieval-based learning (indeed, it is not clear that there is a single mechanism underlying all retrieval-based learning), but focus on the benefits of recall practice considering that test practice benefits are largest after taking a recall practice test as compared to a recognition practice test (Carpenter & DeLosh, 2006; Glover, 1989). We propose a novel mechanism for the benefits of recall practice based on intra-item learning (i.e., learning about the item, as opposed to strengthening associations between the item and retrieval cues), and we tested predictions of this account by examining recall latencies.

Since the classic work of Ebbinghaus (1913), it has been understood that forgetting curves can advance our understanding of learning and memory. Therefore, the appearance of faster forgetting following restudy as compared to a practice test (without feedback) is particularly noteworthy (Carpenter, Pashler, Wixted, & Vul, 2008; Roediger & Karpicke, 2006b; Toppino & Cohen, 2009; Wheeler, Ewers, & Buonanno, 2003). In many cases, there is a

crossover interaction, with restudy producing better performance in the short-term, whereas test practice produces better performance in the long-run. However, interpretation of this result is complicated because in the absence of feedback, there is no opportunity to learn from the failure to recall. This can be remedied by giving feedback upon failure to recall, resulting in better memory performance after test practice as compared to restudy regardless of the retention interval (Butler, Karpicke, & Roediger, 2007; Carrier & Pashler, 1992; Pashler, Cepeda, Wixted, & Rohrer, 2005; Thomas & McDaniel, 2013). Nevertheless, in terms of isolating the unique benefits of a practice test, the use of feedback confounds the situation because it affords all of the benefits associated with restudy (i.e., re-exposure to the target items, regardless of recall success), as well as the long-term benefits of recall practice. Because we aim to test an account of the learning unique to the act of recall, we focus on the benefits of test practice in the absence of feedback.

Consider in detail the nature of the crossover interaction for the testing effect without feedback: restudy produces higher accuracy than test practice when performance is assessed after a brief retention interval, but the opposite pattern is observed following a long retention interval. Despite appearances, this pattern does not necessarily indicate different forgetting rates, and might instead reflect the level of recall success on the practice test. As Kornell, Bjork and Garcia (2011) pointed out, a practice test without feedback produces a “bifurcated” distribution, with a great deal of learning for the items recalled on the practice test, but no learning for the non-retrieved items. In contrast, *all* restudied items receive an increase in memory strength from restudy. If the strengthening from restudy is less than the strengthening from successful recall practice, this can explain the crossover interaction without requiring different forgetting rates. In support of this account, Jang et al. (2012) administered an initial practice test for all items to

divide them into separate pools of recallable and non-recallable items before additional study or test practice, observing that the advantage of restudy before an immediate final test was almost entirely due to strengthening the non-recallable items. In addition, if recall success on the practice test is very high, a practice test is better than restudy even for an immediate final test (Rowland & DeLosh, 2015). However, the bifurcation model is a descriptive account -- it assumes that successful recall produces more learning than restudy but does not specify *why* this is the case or whether these different levels of learning reflect different processes. The current proposal seeks to build upon the bifurcation model by specifying the learning mechanisms underlying this result. This proposal also builds upon formal models of recall, and the common assumption that recall is a two-stage process.

Two-Stage Retrieval Operations

Many formal process models of memory assume that recall is a two stage process: 1) an initial ‘search’ process isolates a candidate memory using the current context and retrieval cues; and 2) a subsequent ‘recovery’ process extracts the details (e.g., semantic, phonological, or orthographic attributes) of the candidate memory to produce on overt response. A failure to recall could arise from either a failure to find the desired memory, or a failure to recover the details of a memory after locating it. To highlight this conceptual distinction, consider an analogy in which long-term memory is a shipping warehouse and memories are packages in the warehouse. To find a specific package, you need to use the attributes of that package to narrow down your search. Some attributes may work better than others for this search process (e.g., ‘rectangular shape’ may describe the majority of the packages whereas ‘taller than four feet’ may apply to only a handful of packages). Assuming that this search process identifies the desired package,

you cannot specify the contents of that package without opening it up, and packages may differ in their ease of opening (e.g., a package wrapped in duct tape versus a tiny bit of scotch tape).

Search and recovery processes exist in most formal models of recall (e.g., Minerva II: Hintzman, 1984; CLS: Norman & O'Reilly, 2003; SAM: Raaijmakers & Shiffrin, 1981), with this distinction serving to describe differences between recognition performance (which is related to the information that guides search) versus recall (which additionally requires recovery). Although these models assume that recall involves two processes, they do not specify different learning for each process, instead assuming that *any* learning makes it easier to isolate a memory *and* easier to extract its details for recall[1]. In developing an account of the benefits from taking a recall practice test, we consider learning for qualitatively different kinds of associations (e.g., between context and the item, between retrieval cues and the item, and between some features of the item and other features of the item). If learning is a dynamic process in which associations are created or strengthened depending on the temporal order of activation, then different kinds of practice may differentially affect these different associations, selectively boosting the search or recovery processes.

The distinction between what is learned from different kinds of practice is shown in Figure 1, which extends the shipping warehouse analogy to the benefits of restudy versus the benefits of recall practice. As seen in the figure, restudying identifies better attributes for searching for relevant packages (i.e., in the case of free recall, a larger set of correct memories are included in the search set, which increases recall accuracy) whereas recall practice doesn't

[1] At least in the SAM model, this assumption was not necessarily made for a strong theoretical reason, but rather as a simplifying mathematical assumption.

add any memories to the search set, but makes it easier to recover the contents of the memories already in the search set (i.e., recall accuracy is not increased, but it is easier/faster to recall the contents of the memories in the search set). Because our account deviates from prior memory models in proposing that the two retrieval operations are differently affected by different kinds of practice, we adopt new terminology for these two processes: 'primary retrieval' versus 'convergent retrieval'.

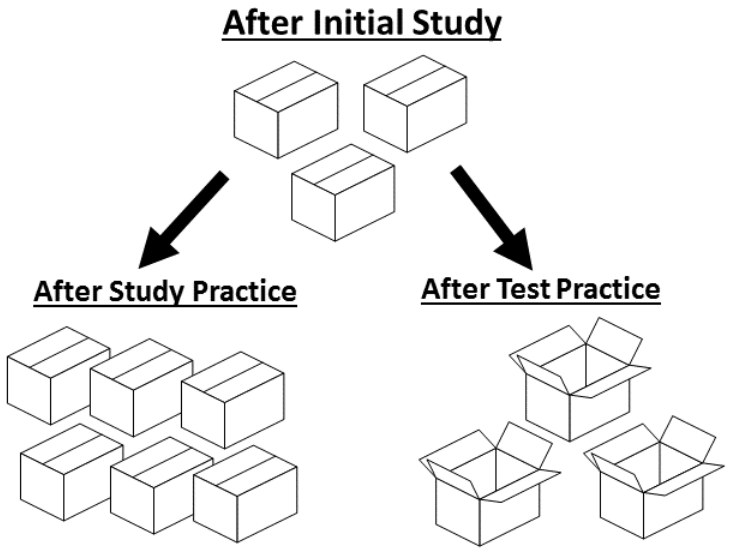

Figure 1: This diagram shows the hypothesized effect of study practice and test practice using a shipping warehouse where packages must be retrieved and opened as an analogy for the memory system. Restudy strengthens the retrieval cues, resulting in a large set of correctly retrieved memory packages (e.g., higher accuracy in free recall or higher familiarity in recognition), whereas test practice makes it easier to reopen memory packages that were successfully recalled during test practice (e.g., higher recall accuracy for previously recalled items as well as faster recall for those items).

Primary and Convergent Retrieval

The Primary and Convergent Retrieval (PCR) model of recall makes 3 core assumptions about how information is recalled:

1) **Primary Retrieval:** In the initial stage of recall, retrieval cues (both context and item cues) activate features of the relevant target memories. Feature activation is likely to be incomplete for any particular item (i.e., some, but not all of the features are active).
2) **Convergent Retrieval:** A subsequent process activates the initially dormant features in one of the items (presumably the most active item). This process may take time to gradually unfold as more and more of the item's features become active. If this process stalls, 'tip of the tongue' occurs (see Brown & McNeill, 1966). However, if this process succeeds, all of the features become active and the item is available for report.
3) **Directional Learning:** Associations between features are directional (e.g., feature A might activate feature B, but not vice versa), and these directional associations are created according to the temporal order in which features become active (e.g., if feature A is active before feature B becomes active, then the directional association from feature A to feature B is strengthened). Because retrieval cues (e.g., context) are active before the presentation of an item for study, study practice results in directional learning from retrieval cues to items. Because successful convergent retrieval is a gradual filling in of an item's features, successful recall (but not study) promotes intra-item learning from some features of an item to other features of an item.

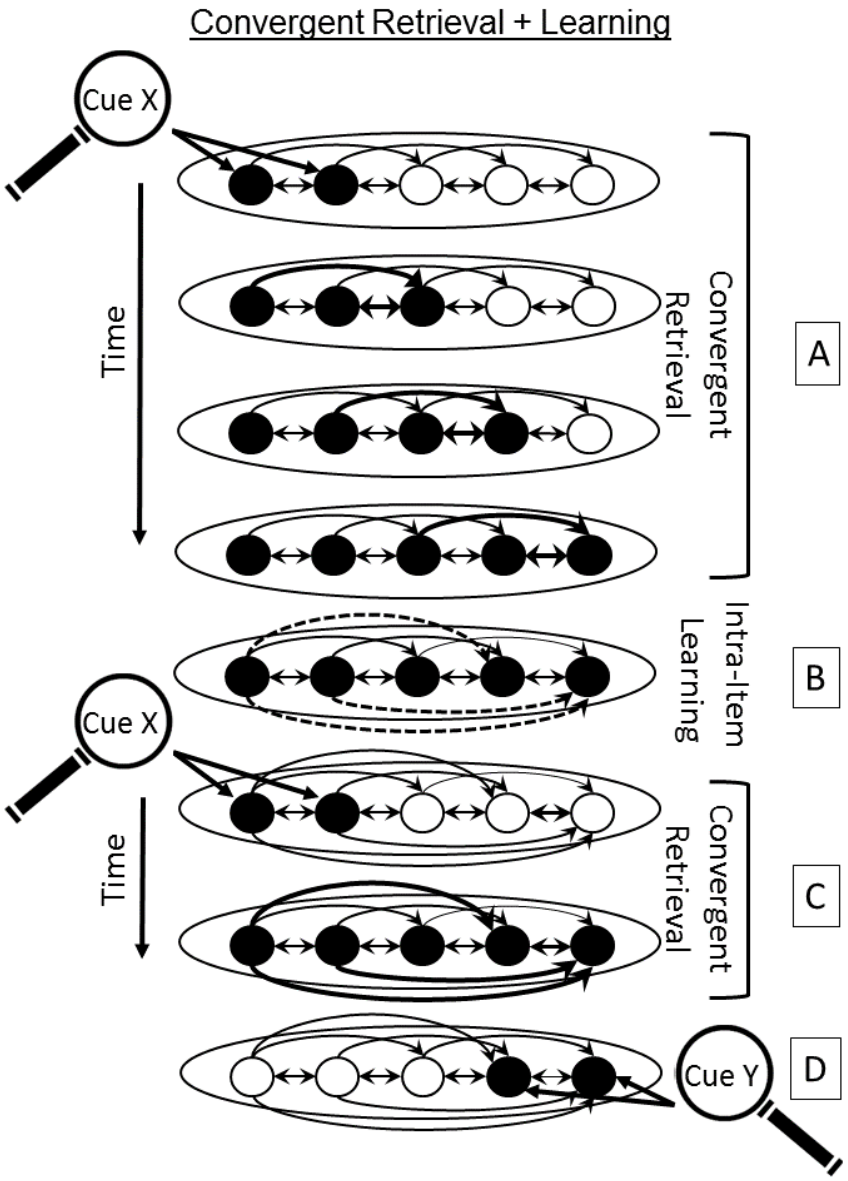

Figure 2: The operations and intra-item learning that occur during convergent retrieval. A) The gradual activation of an item's features during convergent retrieval, initiated by primary retrieval based on retrieval cue X in the first time step. With two lines of support (indicated by the bold arrows), the features of the item are activated one after the other across time steps, resulting in full convergent retrieval and recall success. B) Learning occurs according to the temporal order of feature activations, resulting in new intra-item learning (indicated by the dashed arrows) from features that were active earlier during convergent retrieval to features that became active later during convergent retrieval. C) If retrieval cue X is used for a subsequent recall attempt, convergent retrieval can occur in a single time step owing to intra-item learning from the prior recall success (i.e., decreased recall latency). D) Because learning is directional, initiating convergent retrieval with a different retrieval cue (cue Y), may fail despite prior intra-item learning.

An example of convergent retrieval is shown in Figure 2A based on an item with five features. This is an illustrative example, and a full-fledged version of the model would assign many more features to each item, with features capturing the orthographic, phonemic, semantic, lexical, and perceptual details of the item. In this example, suppose that an inactive feature becomes active if it has two incoming associations from other already active features. The retrieval attempt begins with primary retrieval in the first time step, where the current retrieval cues (cue X) activate the first two features of the item. Feature three then becomes active via its associations from features one and two. Subsequently, feature four is activated via its associations from features two and three. Finally, feature five is activated via its associations from features three and four, and full convergence is achieved. Because convergent retrieval unfolded in a staged manner, new associations are formed between some of the item's features (one → four, one →five, and two → five, as shown by the dashed lines in Figure 2B).

Successful recall is able to produce better retention than restudy because it strengthens the associations between an item's features. If the same features activated in primary retrieval during test practice (or a subset of those features, as occurs after a delay) are activated on a later test, the strengthened intra-item associations make the item more retrievable. In other words, learning directional associations from initially active features to initially dormant features makes it more likely that attempting recall with the original retrieval cues will be successful. More importantly for the current study, intra-item learning reduces convergent retrieval latency (as seen in Figure 2C, a second recall with the same retrieval cues now reaches convergence in a single time step). This prediction of faster retrieval following retrieval echoes the proposal that retrieval latency reflects the number of "decoding" steps required to output an item (MacLeod &

Nelson, 1984). We tested the prediction of faster retrieval following recall practice than restudy in both of the currently reported experiments.

Why should delay result in a subset of the originally activated features? During initial study, context features are active before item features, resulting in directional associations from the context to a subset of an item's features. These new associations provide the basis of primary retrieval. An increase in retention interval is thought to affect the degree of match between the context used for initial learning and the context used at the time of retrieval (Howard & Kahana, 2002; Mensink & Raaijmakers, 1988). Thus, increasing the retention interval changes the retrieval context, and fewer of the originally learned target features will be activated by the changed context. Furthermore, a change of context cannot spontaneously result in the activation of target features that were never associated with the context prior to the delay. Because the features activated by primary retrieval after a context change are a subset of the originally learned features, intra-item learning is critical for convergent retrieval success -- prior recall practice makes it possible to go from this reduced set of initially active features to full convergence. Thus, intra-item learning protects recall performance from forgetting owing to context change.

Besides predicting reduced retrieval latencies and lower forgetting rates, the PCR model predicts that the benefits of recall practice with one previously learned cue will not transfer to a different previously learned cue. For instance, imagine learning two different word cues with the same target word, or learning the target word in two different retrieval contexts, followed by recall practice using one of the two original cues. This is shown in Figure 2D, in which a different retrieval cue (cue Y) was also learned with the item, resulting in a different set of active features after primary retrieval. In this case, recall practice with one cue (cue X), does not help

convergent retrieval based on later retrieval with the other cue (cue Y). In brief, the learning from recall practice is predicted to be cue specific because of directional intra-item learning from initially active features to initially dormant features. We tested this prediction in Experiment 2.

The benefits of recall practice can be contrasted with restudy. Restudy produces new associations between the retrieval cues and the item, enlarging the set of features activated by primary retrieval. This increases the probability of convergent retrieval success (e.g., it is more likely that the augmented starting point will support success) and reduces retrieval latency (having more initially active features reduces the number of features remaining to be activated, thereby speeding retrieval). However, learning from restudy is qualitatively different than learning from test practice because re-presenting the item on a restudy trial activates all its features at once, providing no opportunity for intra-item learning. Over a longer retention interval, the benefits of the enlarged set of features from restudy may be diminished owing to a change in context. To make this concrete, suppose that initial study results in learning 50% of the item's features and restudy results in learning half of the remaining features, such that 75% of the item's features are now associated with the current context. With 75% of the features active in response to the current context on an immediate final test, recall success is likely and will occur quickly (only 25% needs to be filled in). However, after a delay, context is changed, and might, for instance, only contact one third of the originally learned item features. Thus, after a delay, only 25% (reduced from 75%) of the features are activated during primary retrieval, which may be insufficient for recall success given that restudy does not promote intra-item learning (e.g., the links between this 25% of features and the remaining 75% were not strengthened). Furthermore, even if recall success occurs after a delay, it will not be particularly speedy as compared to a situation in which intra-item learning has occurred (e.g., after recall practice, it may be relatively

easy to go from this 25% of features to the remaining 75%). In summary, after a delay, both the initial accuracy benefits and initial latency benefits following restudy are lost as compared to recall practice.

Recall Practice with Free Recall and Cued Recall

Prior studies have shown that recall practice results in faster retrieval on subsequent tests, consistent with the predictions of the PCR model (Keresztes, Kaiser, Kovács, & Racsmány, 2014; Lehman, Smith, & Karpicke, 2014; Pyc & Rawson, 2009; van den Broek, Segers, Takashima, & Verhoeven, 2014; van den Broek, Takashima, Segers, Fernández, & Verhoeven, 2013; Vaughn & Rawson, 2014). For instance, Lehman et al. (2014) had participants learn and practice five different word lists in preparation for a final free recall test covering all five lists. Testing whether context learning underlies retrieval practice benefits, they manipulated the type of practice that occurred after each of the first four lists, including a condition in which participants engaged in recall practice after each list and a control condition without any practice between lists. In all conditions, the critical fifth list was followed by a practice free recall test before the final free recall test. Consistent with the context learning account (and consistent with the context change results of Jang and Huber (2008)), retrieval was faster for the list five practice test when the first four lists were also followed by recall practice. However, this study did not report retrieval latencies from the final test. The current Experiment 1 reports retrieval latencies for both free recall practice and the final free recall test, including a restudy condition to test the prediction that free recall practice should result in faster retrieval latencies on the final test as compared to the situation after restudy.

The study of Van den Broek et al. (2014) assessed cued recall retrieval latencies following recall practice as compared to restudy (also see Macleod and Nelson, 1984). Their study

presented Dutch-Swahili translation word pairs followed by restudy, cued recall practice, or no practice. After practice, the final cued recall test was taken immediately or after a one-week delay. They found that cued recall practice decreased recall latencies more than restudy on both the immediate and delayed final tests, despite producing worse overall accuracy on the immediate test. This finding is consistent with the PCR model, although it should be noted that translation learning may be unique in that it relates the meaning of a known word with a novel word form. In this case, the decreased latency may reflect the learning of the novel orthography/phonology (e.g., learning to speak the translation). Experiment 2 sought to replicate this pattern of results with English word pairs, examining whether these effects generalize to known word forms. In addition, Experiment 2 tested the prediction that these practice effects should fail to transfer between cues.

Figure 3 outlines the possible associative relationships that could be strengthened with test practice. Lehman et al. (2014) suggested that recall practice strengthens associations between the temporal context cues and the item (link 1) and Van den Broek et al. (2014) suggested that cued recall practice strengthens associations between overt retrieval cues and the item (link 2). Associations between the temporal context and the item should be particularly beneficial in free recall testing, where temporal context is the only available retrieval cue (aside from using previously recalled items to cue subsequent retrievals). Associations between overtly provided retrieval cues and the item, such as with word pairs, should be particularly beneficial in cued recall testing, where temporal context is thought to play a diminished role compared to free recall. Assuredly these two types of learning occur to some degree, but the PCR model offers a third possibility; recall practice may strengthen associations between the item and itself (link 3), with these associations resulting in decreased retrieval latencies (as well as increased accuracy in

the case of a delay between practice and the final test). If this account is correct, retrieval latencies should be decreased both with free recall (which primarily depends on temporal context) and with cued recall (which primarily depends on an overtly provided retrieval cue). To test this claim, we performed two experiments, one with free recall and one with cued recall, comparing the accuracy and retrieval latencies after recall practice versus restudy.

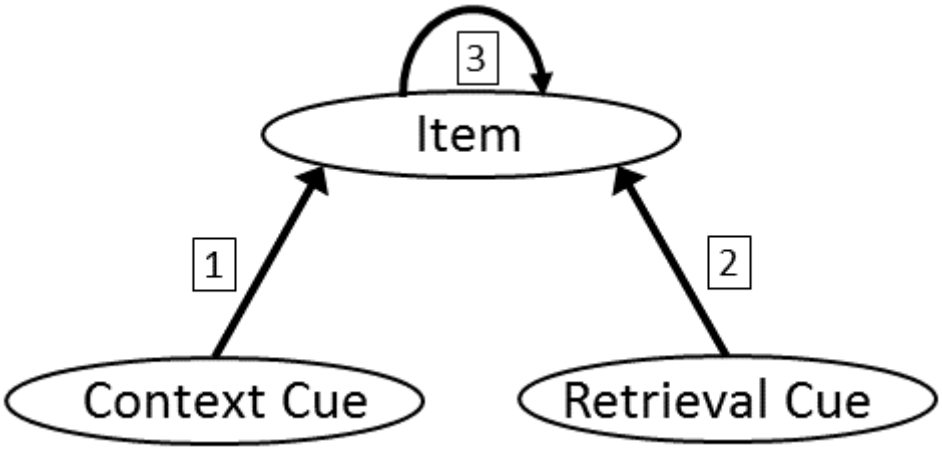

Figure 3: Three kinds of associations that may underlie the benefits of taking a recall practice test, with possible implications for retrieval latencies and accuracy. Associations between the temporal context and the item (link 1) should be particularly beneficial in free recall testing, where temporal context is the only available retrieval cue. Associations between overtly provided retrieval cues (link 2), such as with learning word pairs, should be particularly beneficial in cued recall testing. The primary and convergent retrieval theory suggests a third possibility: that recall practice strengthens self-associations through intra-item learning. On this account, similar results should be found for both free recall and cued recall.

Experiment 1

The PCR model predicts that items successfully recalled on a free recall practice test should be more quickly recalled on a final free recall test as compared to a final test after restudy. PCR makes this prediction because restudy only results in learning from context cues to the item (link 1 in Figure 3) whereas successful recall on a practice test results in this context learning as well as learning between some features of the item and other features of the item (link 3 in Figure 3). This intra-item learning is hypothesized to support convergent retrieval, reducing

retrieval latencies. At the same time, there is should be little or no benefit to accuracy from the free recall practice test provided that the practice test is without feedback. Thus, the PCR model predicts a dissociation between accuracy and retrieval latencies, with restudy increasing accuracy more than a free recall practice test while the practice test decreases retrieval latencies more than restudy.

To test these predictions, participants learned lists of 15 items, followed by a practice phase in which they either restudied the same list or tried to recall as many items as they could from the list. Following the practice phase, participants took an immediate final free recall test on that list. This procedure yields three experimental conditions to compare, baseline (i.e. the practice free recall test), the final free recall test after a restudy opportunity, and the final free recall test after a practice test.

Methods

Participants. 34 individuals from the University of Massachusetts Amherst were recruited from the undergraduate subject pool. Participants were given one unit of credit that could be applied either toward class participation requirements or extra credit opportunities in undergraduate psychology classes. A planned sample size of 30 participants was based on prior literature measuring retrieval latencies (e.g., Roediger & Tulving, 1979; Rohrer & Wixted, 1993, 1994). Recruitment stopped after 30 participants completed the experiment, but already scheduled participants were allowed to complete the study, resulting in a slightly larger sample size. Native language and other demographic information was not collected.

Materials. For each participant, a different random selection of 180 English nouns was made from a pool of 610 words. The word pool was created using from the English Lexicon Project database (Balota et al., 2007). All words in the pool were moderate frequency English

nouns as measured by the SUBTL frequency norms from the $SUBTLEX_{US}$ corpus (Brysbaert & New, 2009) with lengths between three and 10 letters, and concreteness and imageability ratings of over 500 (Wilson, 1988).

Procedure. The experiment used a single factor within-subjects blocked design. Each block consisted of three phases: an initial learning phase, a practice phase, and a final test phase. During the initial learning phase, participants studied a list of 15 serially presented words, where each word was presented alone for three seconds in the center of a computer monitor. Four of the blocks used test practice in the practice phase, while the remaining four used study in the practice phase, with these two types of blocks occurring in alternating order. Whether the first block was test practice versus study practice was counterbalanced across subjects. During study practice blocks, the word list previously studied in the initial learning phase was re-presented to participants in identical serial order, again with three seconds per word. During test practice blocks, participants took a 90 second free recall test where they were instructed to recall as many words as possible from the list of words they had just studied in the initial learning phase. Participants were not permitted to terminate the recall test early, and no feedback was given. During the final test phase of each block, participants took a 90 second free recall test (an identical format as the practice test). Participants typed in responses with the computer keyboard, and could use the backspace key for corrections. Once they were satisfied with their answer, they pressed the enter key to initiate the next recall attempt.

Between initial study and practice, as well as between the practice and final test phases, participants completed a 20 second math distractor task, which involved a running sum of five consecutively presented single digit integers. This design yielded 12 separate memory tests per

subject: four baseline practice tests, four final tests following test practice, and four final tests following restudy. The entire experiment lasted approximately 45 minutes.

Results

Scoring. The accuracy of participants' typed responses was assessed first by an automated routine performing strict string comparison between responses and list items. Recall responses that were scored incorrect by the automated procedure were double-checked and scored by hand, to allow small spelling mistakes to be considered as correct responses. This manual rescoring was performed blind to experimental condition. The reaction time for a specific response in this data set was calculated as the elapsed time between responses (also known as the inter-retrieval time). This elapsed time between each response given was measured by the time between confirming the last correct response[2] (confirmation was given by hitting the "Enter" key after typing in the response) and the first keystroke of the next entered response, except for the first item output, where it was measured as the elapsed time between the onset of the response window and the first keystroke of the first response.

Distractor Task. Participants were 25% accurate on the running summation distractor task. While this performance is somewhat low, the median absolute deviation from the correct answer was only 4, and participants gave a response within the allotted time window on 80% of distractor trials. Thus, they were clearly engaged in this difficult distractor task given that their answers were close to the correct values.

[2] Approximately 3% of responses were errors, consisting of repeats, where the subject typed in a word they had already typed in for a given test period, or intrusions, where the subject typed in a word that was not on the study list. These were removed from the analyses. More specifically, the output position analyses were recalculated as if the error did not occur (i.e., errors did not add an output position) and the inter-response time for a correct response after an error was determined relative to the error (i.e., the subsequent inter-response time did not include the inter-response time of the error).

Recall Accuracy. The proportion of list items that were recalled in each of the three conditions (baseline practice test, final test after restudy, and final test after free recall test practice) were compared with a one-way repeated measures ANOVA. To address violations of sphericity, the degrees of freedom were corrected using Greenhouse-Geisser estimates of epsilon. There was a significant effect of practice type on proportion correct, $F(1.22, 40.17) = 185.03$, MSE = 0.0026, $p < .001$, $\eta^2 = .85$. Bonferroni - corrected *t*-tests showed significant differences in recall accuracy on the final test between the restudy and test practice conditions (81% correct vs 59% correct, $t(33) = 15.11$, $p < .001$, Cohen's $d_z = 1.87$), a significant improvement in accuracy on the final test after restudy relative to the practice test (81% correct vs 61% correct, $t(33) = 12.92$, $p < .001$, Cohen's $d_z = 1.60$) and a small but reliable decrease in accuracy on the final test from the practice test (59% correct vs 61%, $t(33) = -4.68$, $p < .001$, Cohen's $d_z = .18$). Recall accuracy in these three conditions is shown in the left panel of Figure 4.

Recall latency. Free recall inter retrieval times (IRTs) are known to depend on both output position and the total number of items recalled from a list (Murdock & Okada, 1970; Rohrer & Wixted, 1994; Wixted & Rohrer, 1994). Specifically, IRT increases with output position, with each additional recall taking longer (on average) than the last. This is typically a nonlinear accelerating function that builds up to the last output position (i.e., the last few successful recalls are much slower than earlier recalls). As reported by Rohrer (1996), there is a highly consistent pattern to these output position IRT curves as a function of accuracy. He examined natural variation in accuracy under otherwise identical study/test situations (i.e., within condition). If the curves are lined up in the forward output order, accuracy greatly affects the results, and early output position IRTs are faster when more items are ultimately recalled. Remarkably, if the IRT curves are instead lined up in the reverse output order (e.g., comparing

the IRT before the last recall, then the second to last, etc.), the IRT output position curves are nearly identical regardless of accuracy. Thus, "…mean IRT depends on the number of not-yet-recalled items" (p. 195, Rohrer, 1996), which is a unique prediction of the sampling process assumed by many memory models of free recall. However, the key prediction of the PCR model is that IRTs will *also* be affected by the speed of convergence (a.k.a., recovery) and so it is important to use an analysis technique that is uncontaminated by the number of not-yet-recalled items. An analysis by reverse output position achieves this, providing a key test of the claim that test practice produces faster convergence.

Because accuracy was nearly identical for the baseline and test practice conditions, it would not matter whether IRTs were analyzed in the forward versus reverse order directions when comparing these conditions. However, because accuracy was much higher in the study practice condition, an analysis of IRTs as a function of forward output position would be confounded by accuracy, necessarily revealing faster recall at the beginning of the test list because there were more not-yet-recalled items in the restudy condition (for completeness, we report this analysis, finding exactly this result). Thus, to avoid the confound of accuracy, the key analysis of interest considered IRTs in the reverse output position, comparing the time taken to recall the final item in each condition, then the time taken to recall the second to last item in each condition, etc.

Across the entire dataset, the number of items recalled at least once for each condition and each subject was seven, and so this determined the largest possible number of output positions that allowed retention of all subjects in the data analysis. The effects of output position and practice condition on IRTs were assessed with repeated-measures ANOVA. Because the distribution of recall latencies was right-skewed, the raw recall latencies were transformed using

the natural logarithm before calculating a per-subject average, in order to satisfy the ANOVA's normality assumptions. To address violations of sphericity, the degrees of freedom were corrected using Greenhouse-Geisser estimates of epsilon.

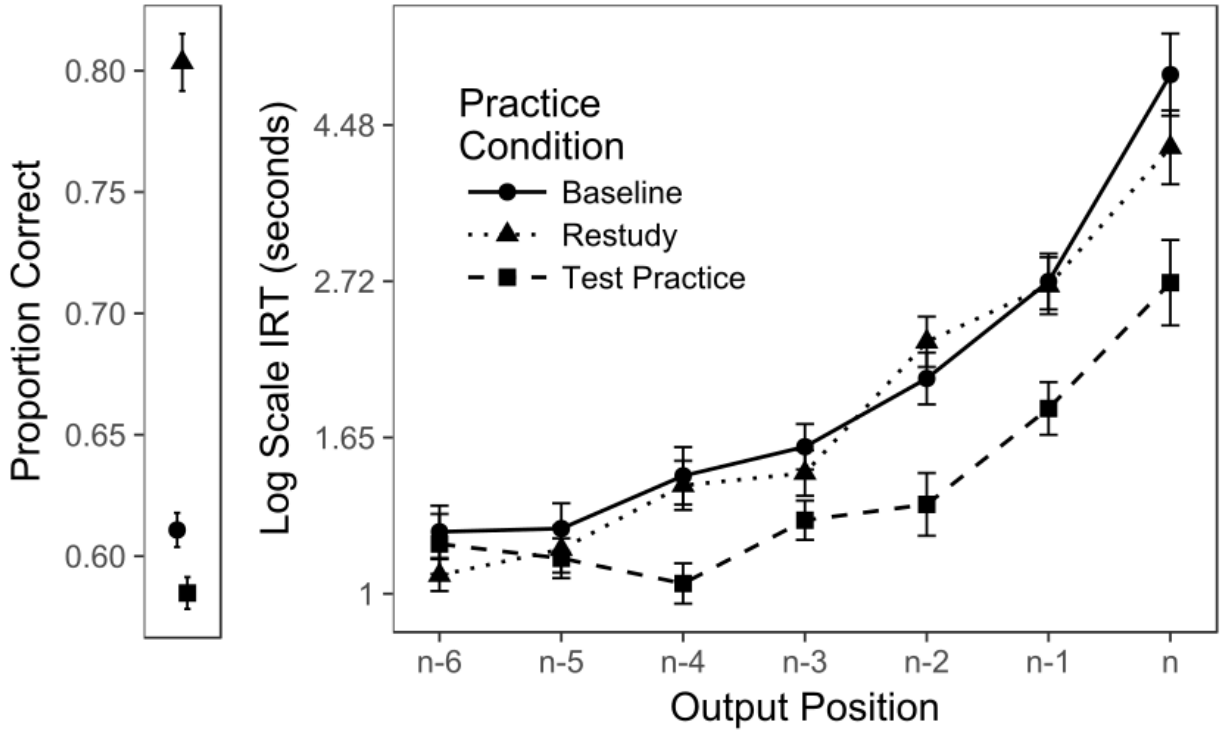

Figure 4: Free recall accuracy (proportion correct, left panel) and Inter Retrieval Times (IRT, right panel) from Experiment 1. Error bars represent +/- 1 SEM calculated using the subject-normalized method of Morey (2008). The "n-th" output position refers to the final item output during a particular recall test.

As explained above, the key analysis considered reverse output position, which, according to prior studies, allows examination of IRTs without accuracy confounding the analysis. The 3 x 7 ANOVA for this analysis revealed a significant main effect of practice type, $F(1.80, 59.29) = 20.53$, MSE = 0.35, $p < .001$, $\eta_p^2 = .38$, a significant main effect of reverse output position, $F(3.39, 111.80) = 59.4$, MSE = 0.61, $p < .001$, $\eta_p^2 = .64$, and a significant interaction between practice type and reverse output position, $F(7.45, 245.73) = 2.93$, MSE = 0.34, $p < .01$, $\eta_p^2 = .08$. As shown in the right panel of Figure 4, recall latencies were fastest in

the test practice condition, and recall latencies in all conditions increased with output position. The nature of the interaction between reverse output position and practice type was investigated further by performing two additional ANOVAs, one considering only the baseline and test practice conditions, and the other one considering only the baseline and restudy conditions. The ANOVA considering only recall latencies from the baseline and test practice conditions revealed a significant difference between the two practice types, $F(1, 33) = 27.85$, MSE = 0.42 $p < .001$, $\eta_p^2 = .46$ and a significant interaction between practice type and reverse output position, $F(4.50, 148.37) = 3.01$, MSE 0.34, $p = .016$, $\eta_p^2 = .08$. On the other hand, the ANOVA considering only the baseline and restudy conditions found no significant differences between the two practice types, $F(1, 33) = 2.05$, MSE = 0.24, $p = .16$, $\eta_p^2 = .06$, and no significant interaction between practice type and reverse output position, $F(4.50, 148.49) = 1.26$, MSE = 0.21, $p = .29$, $\eta_p^2 = .04$). Taken together, these analyses indicate that test practice strongly decreased recall latency from baseline while restudy did not reliably decrease recall latency from baseline, and that the difference from baseline recall latency grew larger over the course of the recall period for test practice items.

For completeness, we also report the analyses based on the forward output order grouping. The 3 x 7 ANOVA revealed a significant main effect of practice type, $F(1.82, 59.96) = 26.10$, MSE = 0.3, $p < .001$, $\eta_p^2 = .44$, a significant main effect of forward output position, $F(4.12, 135.84) = 13.95$, MSE = 0.37, $p < .001$, $\eta_p^2 = .30$, and no interaction between practice type and forward output position, $F(7.05, 232.68) = 1.58$, MSE = 0.29, $p = .14$, $\eta_p^2 = .05$. Differences between the three practice types were assed using Bonferroni-corrected pairwise *t*-tests, using the residual error from the ANOVA in estimating the standard error of the difference. Latencies in the test practice condition were significantly faster than baseline, $t(66) = 2.92$, $p <$

0.05, $d_r = 0.36$, as were the latencies in the restudy condition, $t(66) = 7.18$, $p < 0.001$, $d_r = 0.88$. Latencies in the restudy condition were also faster than in the test practice condition, $t(66) = 4.2$, $p < 0.01$, $d_r = 0.52$, although this is not a meaningful comparison considering the large accuracy differences between these conditions (i.e., when the second item was recalled in the study condition, there were more items yet to be recalled as compared to the time when the second item was recalled in the test practice condition). However, the comparison between the baseline and test practice conditions is meaningful, as these conditions produced similar accuracy levels, and this forward output position analysis produced the same results as the reverse output position analysis, revealing that participants were faster on their second attempt at recalling the same list (even though they were not more accurate).

Discussion

The results of Experiment 1 are in-line with the predictions of the PCR model of the testing effect: Restudy boosted accuracy on the final free recall test, while free recall practice produced slightly lower accuracy on the final test than the baseline test. Despite this lack of improvement following test practice in terms of accuracy, there was a hidden benefit of test practice revealed by examining recall latencies. In general, recall was faster after test practice as compared to the baseline condition. Furthermore, test practice was even faster than restudy when examining recall latencies late in the recall period (e.g., lining up the inter-retrieval times by counting backwards from the last item recalled); prior results established that inter-retrieval latencies should be examined in reverse order to avoid accuracy as a confound (Rohrer, 1996), and this analysis shows an advantage for test practice over restudy, with retrieval latencies following restudy being no different than baseline.

If the same learning processes underlie both restudy and test practice, why should restudy increase accuracy but have no effect on latency while test practice had a negligible effect on accuracy while decreasing latency? This apparent dissociation is explained by the PCR model if restudy primarily affected associations between the temporal context and the items (link 1 in Figure 3 – see also Figure 1) whereas test practice not only affected these associations (but only for items that were successfully recalled during practice) but also the associations between the item and itself (link 3 in Figure 3 – see also Figure 1). In the terminology of the PCR model, restudy boosted primary retrieval for all items (allowing recall of more items) whereas test practice boosted both primary retrieval *and* convergent retrieval, with the latter producing faster recall of previously recalled items. To the best of our knowledge, this is the first demonstration of this dissociation between accuracy and latency when comparing the effects of restudy and test practice in a free recall paradigm.

Lehman et al. (2014) proposed that test practice boosts the association between temporal context and the item (link 1 in Figure 3), but on this account it is not clear why restudy versus test practice produced differing effects when comparing accuracy and latency. Retrieval latencies in free recall are traditionally assumed to reflect the sampling of items from a search set (Rohrer & Wixted, 1994; Wixted & Rohrer, 1994). In other words, the current context activates the list items to different degrees (i.e., the search set) and they compete to be sampled, with each sampling attempt taking some time. If more items from the list are in the search set, this results in higher accuracy, but it takes longer to recall the last few items because of continued resampling of already recalled items. Under any account, restudy must have added more list items to the search set, explaining the observed increase in accuracy. However, with a larger search set, reaction times should have been slower than baseline, and yet restudy produced

similar latencies as compared to baseline. Furthermore, in terms of competition within the search set, it is not clear why retrieval times would be faster after test practice even though accuracy was hardly changed.

To explain these results, the assumption that free recall latencies only reflect a relative competition between items in a search set could be relaxed. Instead, it may be that latency reflects both a relative competition (i.e., the item has to be sampled) as well as the time necessary to recover the item, with the latter determined by absolute retrieval strength (cf. Rohrer, 1996). If test practice increased the absolute strength of recalled items and sped up recovery (such as assumed by the PCR model), this would explain faster retrievals after test practice despite little change to accuracy. It could also explain the lack of latency effects after restudy as reflecting a balancing act, with the increase in latency from an enlarged number of items in the search set being offset by the strengthening of items that would have been in the search set without restudy. However, by proposing that retrieval latencies in free recall are influenced by absolute memory strength rather than just relative memory strength, this account becomes similar to the explanation provided by the PCR model.

In contrast to the proposal of Lehman et al. (2014), Van den Broek et al. (2014) proposed that test practice boosts the association between overtly provided cues and the item (link 2 in Figure 3). However, in the case of free recall, no overt cues are provided. One possibility is that just recalled words are used as cues for the next word to recall, in which case recall practice may have strengthened item associations between these adjacent outputs (e.g., if word B is recalled after word A on the practice test, then on future recall attempts, word A is an effective cue for word B, resulting in faster retrieval of word B). This account predicts that the degree of match

between response order on the baseline practice test and response order on the final test should be closely related to accuracy and reaction time.

To address this possibility, we calculated the Goodman-Kruskal gamma coefficient of output order comparing the practice and final tests for each list, and compared these values to the change in accuracy and recall latency for the corresponding list. The Goodman-Kruskal gamma statistic provides a measure of rank-order correlation based on all pairs of items that were recalled on both the practice test and the final test. According to a response chaining account, larger values of the gamma statistic (i.e., more consistent output ordering between the two tests) should be associated with larger reductions in recall latency between the final and practice tests (i.e., if words are recalled in the same order, this may reflect response chaining, in which case responses should be faster on the final test when the gamma statistic is larger). At the same time, larger values of the gamma statistic should be associated with similar accuracy on the practice and final tests (i.e., if words are recalled in the same order, it follows that a similar number of words would be recalled on both tests). These analyses failed to support an item association account of the results: 1) As seen in the left graph of Figure 5, recalling the same number of words on both tests (the zero point on the x-axis) was not associated with higher gamma values; and 2) As seen in the right graph of Figure 5, the marginal linear relationship between the gamma coefficient and the change in average recall latency for a given list, $r(134) = .149$, $p = .083$, was in the opposite direction from what would be predicted by a response chaining account of the data.

In summary, PCR's prediction of a dissociation between accuracy and latency when comparing the effects of restudy versus a practice free recall test was confirmed. This prediction follows from the proposal that recall success promotes intra-item learning, resulting in faster

recall of the item. These results are incompatible with an item-to-item response chaining account and are only compatible with a temporal context account if one assumes that latencies reflect absolute memory strength in some manner.

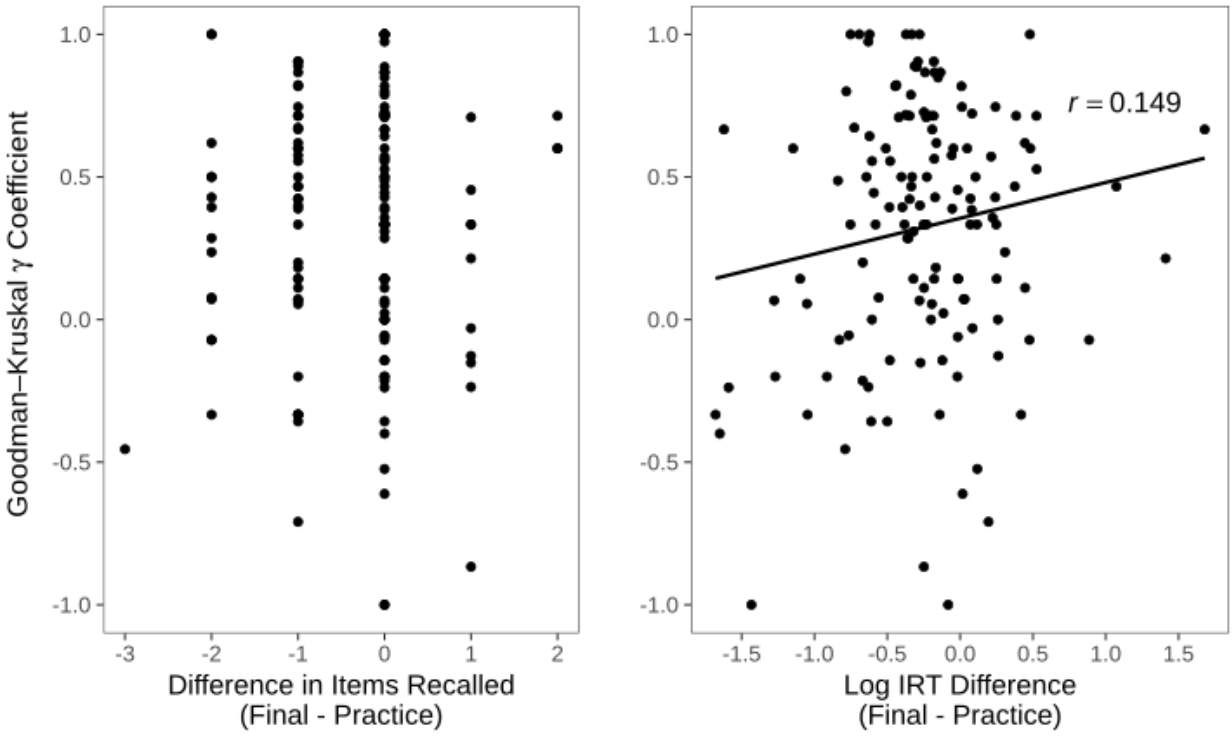

Figure 5: **Left Panel:** Relationship between Goodman-Kruskal gamma coefficient and the change in number of items recalled between final and practice tests for each individual list from Experiment 1. **Right Panel:** Relationship between Goodman-Kruskal gamma coefficient and the change in recall latency between final and practice tests for each individual list from Experiment 1. Solid line represents the line of best fit from regressing the gamma coefficient onto the log-transformed inter-retrieval time differences.

Experiment 2

To better distinguish between the intra-item learning (PCR) and context learning accounts of test practice benefits, Experiment 2 used a cued recall test format with multiple cues and retention intervals. Because Experiment 2 used previously known items (words), context must play a role in retrieval -- for previously known items, the subject's task is to recall based on the episodically defined cue-target association (i.e., learning that occurred in the context of the study list), as opposed to any pre-experimental associations with the cue. Several studies have reported that cued recall practice produces better recall accuracy than restudy (Carpenter et al., 2008; Carrier & Pashler, 1992; Jang, Wixted, Pecher, Zeelenberg, & Huber, 2012; Toppino & Cohen, 2009) and some studies have examined recall latency, finding faster recall after cued recall practice (Pyc & Rawson, 2009; van den Broek et al., 2014; Vaughn & Rawson, 2014). However, these studies did not include conditions for determining whether the test practice benefit was specific to the retrieval cues, which is necessary for identifying the role of context. To identify the role of context, Experiment 2 examined whether the benefits of practice with one cue word transferred to a different cue word in the same context.

In Experiment 2, the initial study list paired a single target item with two different, unrelated cues. This was immediately followed by restudy or cued recall practice with one of these cues, and then a final cued recall test. Final test trials used either the practiced cues (the same cue condition) or the unpracticed cues (the other cue condition). If the benefits of test practice reflect associations between the context and the target, there should be a benefit of test practice even if the final test used the unpracticed cue (provided that the practice test and the final test are contextually similar). Thus, the context learning account predicts the same results (e.g., faster recall, less forgetting) regardless of the overtly provided cues on the final test (i.e., regardless of whether the overtly provided cue is the same as the cue used during practice). If the context

changes between the practice test and the final test, such as might occur with delay, this should weaken the testing benefit for the practiced cue but also the unpracticed cue.

In contrast to a context learning account, the PCR model predicts little or no transfer between cues (no benefit of practicing with one cue on a final test that uses a different cue). This may seem counterintuitive considering that the PCR model proposes that test practice facilitates intra-item learning (learning about the target). However, this follows from the assumption that learning is directional (e.g., learning to recall feature B from feature A does not help recall of feature A from feature B). As a result, the cue plays an important role in the PCR model, setting the initial state of feature activation, producing a cue-specific temporal unfolding of feature activation values during convergent retrieval – one that does not transfer in the event that the final test starts with a different initial state, such as when the final test uses the unpracticed cue.

The PCR model assumes that each item is represented by a collection of features, including semantic features. Words are polysemous, and so the particular collection of semantic features evoked when reading a word is likely to be different on different occasions, depending on the circumstances in which a word is read (e.g., in cue-target word pair learning, the cue word may highlight one specific meaning of the target at the expense of others). For example, reading the target SPEED immediately after reading the cue METER may prompt thoughts of a car's speedometer, which is a particular aspect of speed (i.e., a measure of velocity). In contrast, reading the target SPEED immediately after reading the cue AWAKE may prompt thoughts of the drug amphetamine, which is a different meaning of SPEED. Thus, the evoked meaning of SPEED is different in each case precisely because the two cues are unrelated (METER and AWAKE are randomly chosen cues).

According to the PCR model, directional learning depends on the temporal order of feature activations. For the above example, reading METER then SPEED establishes associations from METER to the features representing the velocity meaning of SPEED whereas reading AWAKE then SPEED establishes associations from AWAKE to the features representing the drug meaning of SPEED. If METER subsequently appeared during test practice and SPEED was recalled, this would strengthen a particular convergent retrieval pathway from velocity-features to the answer SPEED. However, if the final test presented AWAKE, this pathway would not be helpful because AWAKE would not evoke velocity-features. Thus, there would be no test practice transfer between cues. More abstractly, this corresponds to the outcome of the retrieval attempt using Cue X in Figure 2B versus the outcome of the retrieval attempt using Cue Y in Figure 2D. Despite learning from an earlier retrieval using Cue X, Cue Y cannot make use of the strengthened intra-item associations because it is produces a different set of initially active features than Cue X.

Experiment 2 also included an immediate and a delayed final test as a between-subjects manipulation. This delay manipulation is important as it assesses whether the pattern of forgetting is similar between the same cue and other cue conditions. However, unlike previous manipulations of delay, Experiment 2 used a unique design that simulates a long delay in a single session based on the finding that recall promotes context change (Jang and Huber, 2008). For the immediate condition, each initial study and practice test was immediately followed by a final test, whereas in the delayed condition, each initial study and practice test was followed by initial study and practice tests for other lists of words before a final test covering all of the words from all of the lists. Thus, because practice tests on other lists occurred between the initial study/practice and the final test on each list, there was likely to be a context shift.

The PCR model's predictions for the delay manipulation follow from the assumption that delays produce forgetting because the final test context is changed from the original study/practice context (Mensink & Raaijmakers, 1988). On this account, delays can be thought of as a context-switching manipulation, which can be contrasted with the cue-switching manipulation. Critically, the PCR model makes qualitatively different predictions for context-switching versus cue-switching. In brief, context-switching serves to reduce the *quantity* of primary retrieval (e.g., how many features of the target become initially active in response to the test cue) whereas cue-switching serves to change the *quality* of primary retrieval (e.g., which meaning of the target becomes initially active in response to the test cue). In the same-cue condition, although fewer features are activated by the cue, these features will still be the features that were evoked during recall practice, and so prior convergent retrieval practice may allow recall even with a diminished starting point. Returning to the example of practice recalling SPEED from METER, the context change will produce weaker primary retrieval activation of the velocity meaning of SPEED in response to METER, but practice with the convergent pathway from velocity-features to the answer SPEED may protect performance, producing test practice accuracy/latency benefits after a delay. Thus, same cue practice will transfer to other contexts (delay) whereas other cue practice should fail to transfer regardless of delay.

Finally, consider the predictions for restudy. According the PCR model, restudy does not produce intra-item learning and so restudy only strengthens primary retrieval. Thus, after restudy of METER-SPEED, there will be stronger activation of the velocity meaning of SPEED in response to METER on a final test (same cue condition). This stronger activation will make it more likely that convergence will settle upon the correct answer and this stronger activation will also produce latency benefits (it will take fewer time steps to converge because there are fewer inactive features). However, because there was no practice with this specific convergence pathway, the latency benefit from restudy will not be as great as the latency benefit resulting from successful test practice (successful test practice not only results in stronger activation for the velocity meaning of SPEED,

but also practice converging from velocity-features to the correct answer). As with test practice, the PCR model predicts no cue transfer with restudy because each cue-target pair defines a unique primary retrieval starting point.

Methods

Participants. 83 individuals from the University of Massachusetts Amherst were recruited from the undergraduate subject pool. Participants were compensated with one credit that could be applied toward class participation requirements or extra credit points in undergraduate classes. Participants were randomly assigned to either the immediate (n = 39) or delayed (n = 44) final test conditions. A planned sample size of 80 participants (40 in each condition) was based on prior retrieval practice literature that included a between-subjects retention interval manipulation (Carpenter et al., 2008; Jang et al., 2012; Roediger & Karpicke, 2006b; Toppino & Cohen, 2009; Wheeler et al., 2003). We stopped recruitment after 80 participants completed the experiment, but allowed already scheduled individuals to complete the experiment, resulting in a slightly larger sample size. The number of participants tested at each retention interval was unequal because of random assignment to either the immediate or delayed final test condition. We did not collect information about native language or other demographic information.

Materials. For each participant, a different set of 100 word triplets was sampled from the same word pool used in Experiment 1. Word triplets were constructed randomly from the word pool without respect to pre-experimental relatedness. For each triplet, one word was assigned as the target word and the other two words were cue words. Thus, each triplet provided two different word pairs for study, with both word pairs including the same target word. For example, the target word HORSE could appear in two pairs: TABLE–HORSE and STAR–HORSE. Note that randomly chosen cue words will evoke different semantic responses to the target word even

though all three words are nominally unrelated (i.e., they do not appear in the semantic association norms). Thus, TABLE makes you think of a meaning of HORSE in a very different way than STAR does. For instance, a *table-horse* is a kind of bench used for wood working whereas a *horse-star* is the white markings that many horses naturally have between their eyes. The resultant 200 word pairs were organized into 10 lists of 20 pairs each, with each list containing exactly 10 unique target words (i.e., the word pairs with the same target were always assigned to the same list). The presentation order of the word pairs in each list was randomly permuted for each subject.

Procedure. Participants completed three phases for each of the 10 lists: an initial study phase, followed by a practice phase, and ending with a final test phase. In the initial learning phase, participants studied each of the 20 word pairs in the list, one pair at a time, on a computer screen for four seconds each. Cue and target words were presented on the left and right sides of the screen, respectively. In the practice phase, four of the target words were randomly selected to receive restudy, four of the target words were randomly selected to receive cued recall test practice, and two of the target words were not practiced. The targets selected to receive additional practice were only practiced with one of the two cue words they were studied with during the initial learning phase. For example, if HORSE was a target selected to receive restudy, then either TABLE–HORSE *or* STAR–HORSE would be restudied, but not both. Likewise, if HORSE was a target selected to receive cued recall test practice, then participants would either be shown either TABLE *or* STAR as a cue to recall the word HORSE.

The restudy and test portions of the practice phase were blocked, and the order of the study/test blocks within the practice phase was counterbalanced across lists. On restudy trials during the practice phase, word pairs were displayed for four seconds. On cued recall trials,

participants were shown the cue word on the left side of the screen, and a question mark prompt on the right side of the screen. Participants were given up to 8 seconds to initiate a response using the computer's keyboard. Provided that the response was initiated within 8 seconds, they could take as long as needed to complete the typing of their response. Participants were permitted to edit their responses using the Backspace key before confirming them with the Enter key. No feedback was given on the practice test. After typing Enter, the next practice test trial began immediately, which served to limit the amount of time available for dwelling upon a correctly recalled target. This was done to highlight the differences between restudy and test practice (i.e., most of the time in test practice was spent retrieving rather than reviewing). Although they could take up to 8 seconds to initiate a response plus the time needed to type an answer, subjects typically responded much more quickly, and the average time from the start of a test practice trial until the initiation of the next trial (i.e., total time for retrieval and typing) was only 4.37 seconds. Thus, the average total time spent on test practice trials (4.37 seconds) was comparable the total time spent on a study practice trials (4 seconds). The Two alternating rounds of restudy and test practice were given for each list (e.g., Study – Test – Study – Test). The relative order of word pairs within each practice block was the same as in the initial study phase, but because word pairs were randomly assigned to a practice condition and the order of the study/test practice blocks was randomly chosen, the absolute order of the word pairs was different than in the initial study phase.

After the practice phase, each target was given a final cued recall test using only one of its associated cue words from the initial study phase. Half of the targets given restudy were tested using the same cue word that was used during restudy (the same-cue restudy condition), while the other half were tested using the unpracticed cue (the other-cue restudy condition). This was

done on a list-by-list basis, such that each list contained two same-cue restudy trials and two other-cue restudy trials. Similarly, half of the targets given cued recall practice were tested with the same cue word that was used during the practice test (the same-cue test condition), while the other half were tested using the unpracticed cue (the other-cue test condition). Target words that were not practiced were tested with one of their cues from the initial study phase (the baseline condition). As during the practice tests, participants were given 8 seconds on test trials to type in the missing target word using the computer's keyboard. The order of the word pairs was randomly shuffled within each list, so that the order of pairs on the final test would not be the same as in the study or practice phases.

The timing of the final test was different for the immediate and delayed final test conditions. Participants in the immediate final test condition completed the final test of a list immediately after practice for that list. Thus, they experienced ten rounds of initial study, practice, and final test. For the delayed final test condition, the final test for all ten lists did not occur until after all ten lists had received initial study and practice. Thus, they experienced ten rounds of initial study and practice, and then one long final test covering the targets from all ten lists (e.g., study-practice for lists one, two, three, etc., followed by the final test for list one, then the final test for list two, etc.). To maintain consistency with the immediate condition, the lists were tested on the final test in the same order they were studied in. A 10 second break was given in between the final test for each list (i.e., after every 10 trials).

Results

Participants cued recall responses were scored for accuracy in the same manner as Experiment 1. Recall latency for each response was measured as the duration between the presentation of the retrieval cue, and the first keypress of the participant's response. The degrees

of freedom in all *F*-tests involving repeated measures factors were corrected using Greenhouse-Geisser estimates of epsilon to account for violations of sphericity.

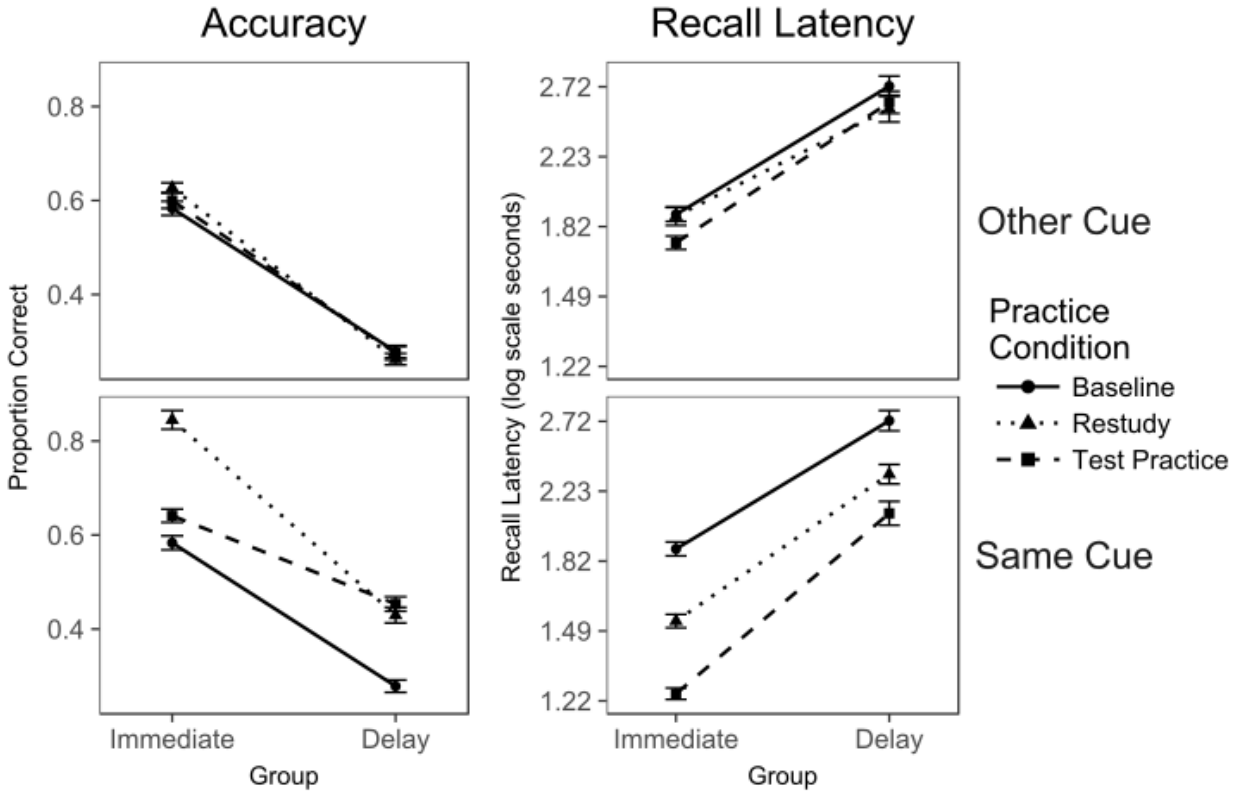

Figure 6: Cued recall memory accuracy and recall latency results from Experiment 2. Error bars represent +/- 1 SEM calculated using the subject-normalized method of Morey (2008). Note that the baseline condition is duplicated across both rows of the figure in order to enable easier comparison between baseline and each experimental condition.

Recall Accuracy. The proportion of list items that were recalled in each of the five within-subject conditions (baseline, same-cue restudy, other-cue restudy, same-cue test, and other cue test) and the two between-subject conditions (immediate and delayed final test) are shown in the left column of Figure 6, and were statistically compared with a 5 by 2 mixed ANOVA. The ANOVA revealed significant main effects of retention interval, $F(1, 81) = 50.69$, MSE = 0.21, $p < .001$, $\eta_p^2 = .38$, practice type, $F(3.54, 286.4) = 73.11$, MSE = .01, $p < .001$, $\eta_p^2 = .47$, and a significant interaction between retention interval and practice type, $F(3.54, 286.4) = 16.07$, MSE

= 0.01, $p < .001$, $\eta_p^2 = .17$. As expected, memory accuracy was higher in the immediate final test condition than in the delayed final test condition. At both retention intervals, items in the same-cue restudy and same-cue test conditions were recalled more accurately than baseline items, while accuracy in the other-cue restudy and other-cue test conditions was not different from baseline. However, the relationship between the same-cue restudy and same-cue test conditions reversed across the two retention intervals; accuracy for same-cue restudy items was greater than same-cue test items in the immediate final test condition, while accuracy for same-cue test items was greater than same-cue restudy items in the delayed final test condition. This cross-over in performance replicates the classic testing effect when using a delay of several days, indicating that interleaving lists to promote context change was effective at producing results equivalent to a long delay within a single session.

Recall latency. The recall latencies for correctly recalled items in each condition are shown in the right column of Figure 6, and were statistically compared with a five by two mixed ANOVA. Because the distribution of recall latencies was right-skewed, the recall latencies were transformed using the natural logarithm before analysis, in order to meet the ANOVA's statistical assumption of a Gaussian random variable. Additionally, six subjects (one from the immediate final test condition, and five from the delayed final test condition) were excluded from this analysis due to missing observations from at least one of the practice type conditions (i.e., they failed to recall any of the items from at least one condition). The ANOVA revealed significant main effects of retention interval, $F(1, 75) = 62.09$, MSE = 0.23, $p < .001$, $\eta_p^2 = .45$, practice type, $F(3.63, 271.93) = 50.66$, MSE = .03, $p < .001$, $\eta_p^2 = .4$, and a significant interaction between retention interval and practice type, $F(3.63, 271.93) = 3.90$, MSE = .03, $p = .006$, $\eta_p^2 = .05$. Items were recalled more quickly (i.e., lower recall latencies) in the immediate final test

condition than in the delayed final, across all practice types. At both retention intervals, items in the same-cue restudy and same-cue test conditions were recalled faster than baseline items, while items in the other-cue restudy and other-cue test conditions had recall latencies similar to baseline items. However, the presence of the significant interaction between retention interval and practice condition indicates that the recall latency differences between practice types was not uniform over both retention intervals.

To further investigate the interaction between retention interval and practice type, we performed separate two by three mixed ANOVAs for same-cue and other-cue items (baseline items were included in each ANOVA). In these ANOVAs, the main effect of practice type and the interaction between practice type and retention interval are the critical comparisons. The ANOVA comparing same-cue restudy, same-cue test, and baseline items across both retention intervals found a significant main effect of practice type, $F(1.90, 142.25) = 83.17$, MSE =0.03, $p < .001$, $\eta_p^2 = .53$, and a significant interaction between retention interval and practice type, $F(1.90, 142.25) = 4.47$, MSE = 0.03, $p = .014$ $\eta_p^2 = .06$. In the immediate final test condition, same-cue test items were recalled faster than same-cue restudy items, which were in turn recalled faster than baseline items. The difference between same-cue test and same-cue restudy items decreased in the delayed final test condition, as did the difference between the same-cue restudy and baseline items, producing the interaction effect. The ANOVA comparing the other-cue restudy, other-cue test, and baseline items across both retention intervals found no main effect of practice type, $F(1.93, 144.72) = 1.7$, MSE = .03, $p = .19$, $\eta_p^2 = 02$, and no interaction between retention interval and practice type $F(1.93, 144.72) = 1.74$, MSE = .03, $p = .18$, $\eta_p^2 = 02$. This indicates that the same-cue items were the drivers of the practice type main effect and practice type by retention interval interaction effects seen in the full two by five ANOVA.

Discussion

The results of Experiment 2 provide further support for the intra-item learning account of testing effects. As predicted, cued recall practice produced faster recall latencies on both the immediate and delayed final tests than restudy, which in turn produced faster recall latencies than no practice. The recall latency advantage for cued recall practice over restudy was consistent across retention intervals, even though there was a crossover interaction in terms of accuracy. This replicates previous findings (van den Broek et al., 2014), but with pairs of well-known English words (i.e., learning to relate the meaning of two different words), in contrast to translation learning (i.e., learning to relate the meaning of a known word with a novel word form). In addition, this experiment builds on previous results by comparing same cue practice results with a cue-switching condition (other cue practice). Critically, the same/other cue manipulation was within subject and within list, and yet practice with the other cue failed to produce any advantages either in terms of accuracy or latency. This rules out a context learning account of testing effects with cued recall, as such an account cannot simultaneously explain the benefits of test practice in the same cue condition, and the lack of benefits in the other cue condition. Under a context learning account, the context of the practice session must have been sufficiently similar to the final test so as to afford a learning advantage in the same cue condition. Thus, if same cue practice produced a context learning advantage, the other cue practice should have likewise produced a context learning advantage. The results of this experiment are compatible with a cue learning account (i.e., link 2 in Figure 3), explaining why practice with one cue failed to transfer to the other cue. However, a cue learning account cannot explain the results of Experiment 1, which did not provide any cues for recall. In contrast to these alternative explanations (i.e., links 1 and 2 in Figure 3), the intra-item learning account proposed in the PCR

model (i.e., link 3 in Figure 3) provides a consistent explanation of the similar results found with free and cued recall.

Based on the PCR model's assumptions of feature representations and directional associations, the model predicted that practice with one cue will fail to transfer to another cue. More specifically, because each cue was learned on separate study trials during initial study, and because the cues were unrelated to each other and to the target, the primary retrieval starting points for each cue were expected to be different (i.e., each cue was associated with a different set of target item features, such as with the METER-SPEED = speedometer and velocity features whereas AWAKE-SPEED = amphetamine and drug related features example). According to the PCR model, successful recall practice using the cue provided during test practice strengthens a convergent retrieval pathway specific to the primary retrieval staring point provided by that cue. If memory for that item is probed with that same cue in the future, the strengthened intra-item associations will benefit the convergent retrieval process (i.e., recall is faster). These intra-item associations protect the item from forgetting effects with delay; with a context change after a delay, a smaller number of features will be activated by the cue, but, nevertheless, owing to intra-item associations from these features, convergent retrieval may still be possible. However, if memory for that item is probed with an unpracticed cue (one that was also learned during initial study), then the intra-item associations learned from the starting point of the practiced cue may fail to benefit the convergent retrieval process because that process starts in a different place. Furthermore, this lack of transfer should occur regardless of delay.

The lack of transfer between cues in Experiment 2 may appear at odds with previous reports of retrieval practice transfer between retrieval cues and generalization from retrieval-based learning (Butler, 2010; Carpenter, Pashler, & Vul, 2006; McDaniel, Anderson, Derbish, &

Morrisette, 2007; Rawson, Dunlosky, & Sciartelli, 2013; Vaughn & Rawson, 2014), though transfer effects after test practice are not universally found (Hinze & Wiley, 2011; Pan, Gopal, & Rickard, 2016; Pan, Wong, Potter, Mejia, & Rickard, 2015). However, prior studies reporting transfer effects examined situations where the cues were related to the target information. According to the PCR model, transfer is expected with related cues to the extent that both cues evoke a similar primary retrieval state (e.g., BRAKE-SPEED and DRIVE-SPEED will both evoke car-related speed information and so practice with BRAKE-? will benefit DRIVE-?). In contrast, the other cue condition of Experiment 2 examined transfer between cues that were unrelated to each other and to the target. Because of these differences, Experiment 2 is unique in the recall practice literature, and the other cue condition demonstrates an important caveat for benefits of from testing.

General Discussion

The Primary and Convergent Retrieval (PCR) model of recall builds on the assumption made by most memory models that recall is a two-stage process, with an initial stage (primary retrieval) using context and any other cues to specify a search set of possible memories, followed by a second process (convergent retrieval), which attempts to fill in any missing pieces (i.e., pattern completion) for a specific memory within the search set, with full convergence being necessary for overt production of the item. However, unlike previous memory models, the PCR model proposes learning processes that are unique to this second stage of recall. More specifically, by adopting a feature-based representation of items in memory and by assuming that feature-to-feature associations are directional and learned from the temporal order in which features become active, the PCR model predicts that the act of successfully recalling an item (i.e., successful convergence) will strengthen associations between the features of that item. This

intra-item learning does not occur with study of an item because the item's features are presented all at once with study. Because this intra-item learning is about the item, rather than the association between context and item, it predicts that the benefits of recall practice will reduce the rate of forgetting, to the extent that forgetting occurs because of context change.

The PCR model specifies the dynamic time course of recall; specifically, the convergent retrieval process is assumed to be an important factor in determining recall latency. As a result of intra-item learning, not only is it more likely that the item will be recalled on future memory tests, but it should take less time to recall the item (i.e., convergence in fewer time steps). We confirmed this prediction with two experiments that measured both recall latency and memory accuracy as a function of whether information was practiced with testing, restudy, or not practiced. Across both experiments, recall latencies were fastest following recall practice for an immediate final test. Thus, as predicted by the PCR model, there was a dissociation between accuracy and latency when comparing the effects of restudy versus a practice recall test, as seen in Figure 7, which shows a state-trace analysis of this dissociation (Bamber, 1979) by plotting average recall latency and memory accuracy against one another for the immediate final tests from Experiments 1 (free recall) and 2 (cued recall). This shows that restudy primarily increased memory accuracy relative to baseline, whereas test practice primarily decreased retrieval latency. In addition to this dissociation between recall and accuracy for an immediate final test, Experiment 2 found a dissociation between these measures as a function of delay. More specifically, there was a latency advantage regardless of delay (test practice faster than restudy) even though there was a crossover interaction between test practice and restudy as a function of delay when considering accuracy.

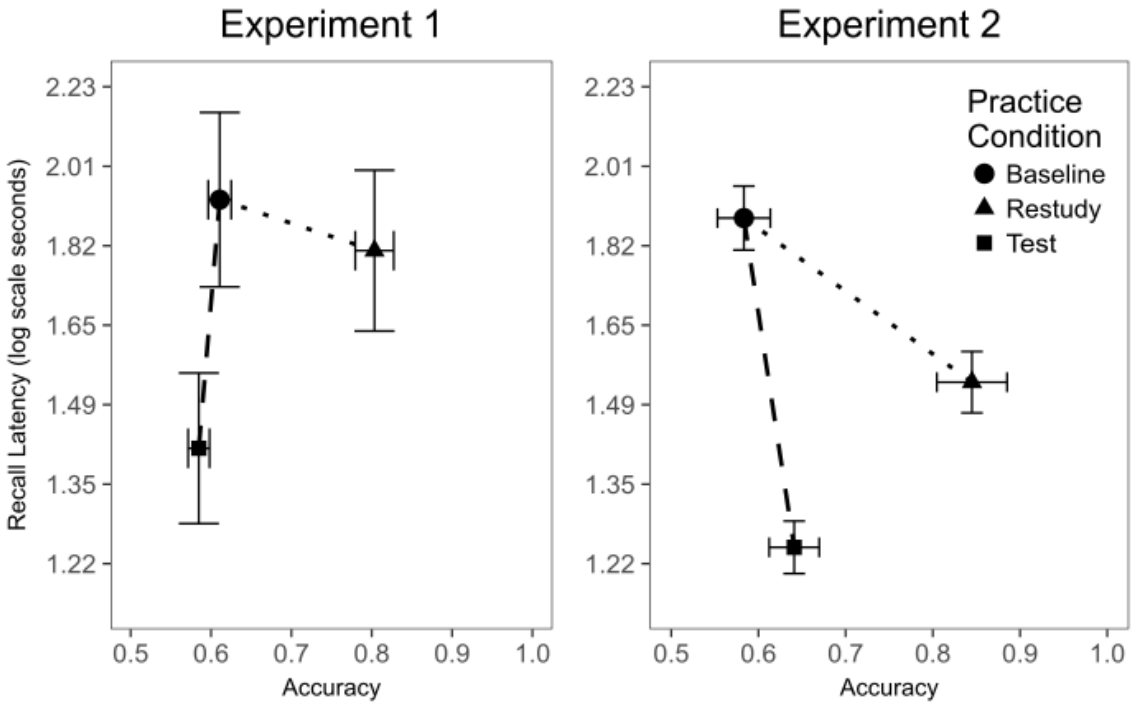

Figure 7: Average accuracy and recall latency results from the immediate final tests in Experiments 1 and 2. Restudy has the primary effect of increasing accuracy, while recall practice has the primary effect of reducing recall latency. Only the results from the same-cue conditions in Experiment 2 are shown. Error bars represent 95% confidence intervals calculated using the subject-normalized method of Morey (2008).

Roughly speaking, there are three possibly types of learning from the act of taking a recall test: Strengthened associations between the temporal context and the item, strengthened associations between overtly provided retrieval cues and the item, and strengthened associations between the item and itself. It is this third possibility that is the novel contribution of the PCR model. Furthermore, with its feature-based representation and directional associations, the PCR model predicted that intra-item learning should be cue-specific because unrelated retrieval cues specify independent sets of learned item features. By examining the pattern of results across both free recall and cued recall, and by examining whether recall practice with one cue transfers to a different cue, we found support for intra-item learning.

In theory, strengthened associations between cues and target memories can explain free recall testing effects if test takers learn to use previously recalled items as cues for subsequently recalled items. However, our response order analyses of the free recall data failed to find support for this cue-learning account. In contrast, context learning can potentially explain the free recall data. However, context learning cannot explain why taking a practice test with the same cue as used on a final test boosts performance (decreasing latency regardless of delay), whereas a practice test with a different cue than the one used on a final test failed to produce any effects in terms of accuracy or latency. It is possible that different learning mechanisms underlie the benefits of taking a free recall practice test as compared to a cued recall practice test, but intra-item learning provides a parsimonious account of both forms of recall practice.

Theoretical Accounts of Retrieval Practice Learning

The PCR model is far from the first explanation of learning from recall practice. In this section, we discuss its relationships with other theories.

Transfer-appropriate Processing. One of the oldest explanations of learning from recall practice is transfer-appropriate processing (Morris, Bransford, & Franks, 1977). Transfer-appropriate processing is a general learning principle, stating that performance will be better to the extent that the processes recruited during learning are the same as the processes necessary on a later test. This principle explains why a recall practice test is more effective than restudy in preparation for a later recall test. Despite its intuitive appeal, transfer-appropriate processing is somewhat descriptive, failing to indicate the nature of the processes involved in retrieval. Furthermore, systematic comparisons between different kinds of practice and different kinds of final tests failed to support transfer-appropriate processing as an all-encompassing explanation for the benefits of practice tests (Carpenter and DeLosh, 2006; also see Glover, 1989).

Nevertheless, the principle of transfer-appropriate processing assuredly applies in many situations, and the PCR model can be seen as a specific model implementation of transfer-appropriate processing by proposing that the act of successful convergent retrieval lends itself (via intra-item learning) to subsequent convergent retrieval success.

Effortful Retrieval. Similar to transfer-appropriate processing, the theory of effortful retrieval is also a general learning principle. This principles states that the degree of learning from a practice test is determined by the difficulty of retrieval, with difficult but ultimately successful retrieval producing greater learning (Bjork, 1975). In general, this principle is well supported in the literature on testing effects. For example, Carpenter and DeLosh (2006) found that free recall practice tests produced the best final test performance, regardless of final test format (in contradiction to transfer-appropriate processing). This result follows from the principle of effortful retrieval because free recall is more difficult/effortful than cued recall or recognition (i.e., cued recall and recognition provide more cues to aid retrieval). Other studies have manipulated the spacing between initial encoding and practice tests, seeking to make the practice tests more difficult but nevertheless successful (Karpicke & Roediger, 2007; Pyc & Rawson, 2009). As predicted by the principle of effortful retrieval, these studies found that longer retention intervals between initial encoding and the practice test enhanced the magnitude of the testing effect. A meta-analysis of testing effect studies reported evidence for retrieval effort as a moderator of the testing effect, due in large part to the greater magnitude of testing effects when the practice test uses a relatively difficult format, such as free recall, as opposed to a recognition practice test (Rowland, 2014).

As with transfer-appropriate processing, the principle of effortful retrieval is somewhat descriptive, failing to specify why greater effort results in more learning. Bjork and Bjork's

(1992) theory of disuse represents one possible model instantiation of a retrieval effort theory. Under this theoretical account, an item's memory strength is multifaceted: memories have separate *retrieval* strength (representing the memory's current accessibility) as well as a *storage* strength (representing the degree to which the item is well-learned or engrained in memory). A memory's current retrieval strength determines the probability of recalling the memory whereas a memory's storage strength moderates changes to its retrieval strength. More specifically, higher storage strength potentiates increases in retrieval strength (learning), and slows the decline of retrieval strength over time (forgetting). Studying and successful retrieval are thought to increment both an item's retrieval and storage strength. However, for two items with equal storage strength but different retrieval strengths, the item that with a lower retrieval strength receives a greater increment to its retrieval strength as a result of successful retrieval. Thus, the theory has an account of why greater learning occurs from difficult retrieval: an item that is difficult to recall is one with low retrieval strength, which in turn allows for greater learning.

The PCR model shares several characteristics with the theory of disuse. Like the theory of disuse, the PCR model assumes that an item's memory strength is multifaceted, including both primary retrieval (i.e., the quantity and quality of its associations with the current retrieval cues, which is analogous to retrieval strength) and convergent retrieval (i.e., the quantity and quality of the associations between the features of the item, which is analogous to storage strength). Also, similar to the theory of disuse's assumption that storage and retrieval strength can be separately altered, the PCR model assumes separate learning for primary retrieval and convergent retrieval. Although the PCR model does not define 'difficulty' or 'effort', it is reasonable to assume that a convergent retrieval process taking more time steps will give rise to a phenomenological experience of greater effort/difficulty. With more time steps to convergence, it follows from

PCR's learning assumptions that more intra-item learning occurs; there will be more specific pairwise instances of one feature being active before another. According to the PCR model, this multi-step effortful retrieval is more likely to occur for items with initially poor intra-item associations (similar to the theory of disuse's assumption of greater learning for items with initially low retrieval strength). From this perspective, the PCR model could be viewed as a detailed instantiation of the theory of disuse by specifying the feature-to-feature learning and retrieval processes that underlie difficulty and different kinds of memory strengths. By considering these processes in greater detail, the PCR model makes specific predictions regarding recall latencies (in the theory of disuses, it is unclear which memory strength maps onto latency).

Elaborative Retrieval. The elaborative retrieval hypothesis holds that retrieval enhances subsequent memory because it affords the opportunity to *elaborate* on the relationship between the current retrieval cues and the target item in memory (Carpenter, 2009, 2011; Carpenter & Yeung, 2017). Specifically, the theory proposes that during testing, participants activate cue-relevant information (e.g., semantic associates of the retrieval cues and the target word), and that activation of this information is beneficial because it enhances later access to the target item. Restudy and less effortful test practice formats (e.g., recognition) do not induce semantic elaboration, so the benefits from these practice methods are less pronounced.

The PCR model and the elaborative retrieval hypothesis are similar in some ways, but critically differ in other ways. One on hand, both theories propose that recall practice provides the opportunity to enhance a retrieval pathway that is not used with restudy, and both theories hold than this can involve semantic information relating the cue and target. On the other hand, they differ as to the locus of learning that leads to recall practice benefits. According to the

elaborative retrieval hypothesis, the retrieval pathway unique to recall practice is through associations with other distinct items in memory whereas the PCR model assumes that the pathway unique to recall practice is between different features within the item.

To date, the elaborative retrieval hypothesis has not been applied to recall latencies. Intuitively, it might seem that retrieval via associations with other items in memory would be slower rather than a faster. However, this only follows if retrieval is a serial process, going from retrieval cues to other items in memory, and finally to the desired target item. If retrieval is instead a parallel process, these elaborated associations with other items in memory may provide a collaborative boost, with rapid convergence on the target. The proposal that test practice with specific cues promotes retrieval paths via semantic associates is compatible with the results from the other-cue conditions in Experiment 2, which failed to produce transfer effects. However, other studies have pointed out problems with the elaborative retrieval hypothesis. For instance, having participants overtly generate associates in response to a retrieval cue reduces accessibility of a particular target rather than making retrieval easier (Watkins & Watkins, 1975). Furthermore, attempts to measure and induce elaboration during practice have failed to find a positive relationship between the amount of elaboration and subsequent retention (Lehman & Karpicke, 2016).

Retrieved Context Account. As discussed previously, a recently proposed explanation of learning from retrieval practice appeals to the updating of stored contextual representations during the practice test (Karpicke, Lehman, & Aue, 2014; Lehman et al., 2014). Because retrieved items are updated to reflect the current context of the practice test, they are stored with a context that is more similar to the context of the final test. If this updating does not occur with restudy, or perhaps occurs to a lesser extent, this explains the long-term benefits of test practice

over restudy. This account is well-integrated with established theories of memory retrieval, building upon retrieved context models, which successfully explain the organizational patterns of free recall behavior (Howard & Kahana, 2002; Lehman & Malmberg, 2013; Polyn, Norman, & Kahana, 2009).

The retrieved context account and the PCR model fundamentally differ in their assumed learning mechanisms. Under the retrieved context account, retrieval practice results in a better match between the target memory and the context cues used on the final test. Within the PCR model, this is akin to enhanced primary retrieval (i.e., associations between context and item features) rather than convergent retrieval (i.e., association between item features and other item features). The retrieved context account predicts faster retrieval latencies after successful test practice, but for different reasons than the PCR model. Under the retrieved context account, retrieval latencies are reduced because the item is more likely to be sampled within the search set of contextually appropriate memories (Lehman, Smith, & Karpicke, 2014; Rohrer & Wixted, 1994b, 1993). In contrast, the PCR model assumes that latencies are reduced because of a change in the recovery process (which is relabeled convergent retrieval in PCR) rather than the sampling process (which is relabeled primary retrieval in PCR). Comparisons of these accounts based on free recall latencies will require further investigation (e.g., consideration of latency distributions, manipulations of list-length, etc.) to determine whether test practice primarily influences the sampling process or the recovery process. Nevertheless, in the current study, we compared these accounts in a different way by considering the benefits of cued recall practice, finding that practice with the same cue as the final test produced a latency benefit whereas no benefits were found for cued recall practice with a different cue than the one used on the final test. This lack of transfer is difficult to reconcile with the retrieved context account unless the learning mechanism

underlying the benefits of cued recall practice is different than the learning mechanism underlying the benefits of free recall practice.

Final Conclusions

The Primary and Convergent Retrieval (PCR) model is a novel theoretical account for the learning benefits from taking a practice recall test. Specifically, the PCR model assumes that recall practice causes greater learning than restudy by strengthening associations between the features of the item. We confirmed predictions arising from this intra-item learning account, finding dissociations between recall accuracy and recall latency for both free recall and cued recall practice tests. These results demonstrate that accuracy and latency need to be jointly considered when evaluating different theories of learning as they relate to testing effects and effective study.

References

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology*, *19*(2), 137–181. https://doi.org/10.1016/0022-2496(79)90016-6

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. F. Healy, S. M. Kosslyn, R. M. Shiffrin, A. F. (Ed) Healy, S. M. (Ed) Kosslyn, & R. M. (Ed) Shiffrin (Eds.), *Essays in honor of William K. Estes, Vol. 1: From learning theory to connectionist theory; Vol. 2: From learning processes to cognitive processes.* (pp. 35–67). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Brown, R., & McNeill, D. (1966). The “tip of the tongue” phenomenon. *Journal of Verbal Learning & Verbal Behavior*, *5*(4), 325–337. https://doi.org/10.1016/S0022-5371(66)80040-3

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(5), 1118–1133. https://doi.org/10.1037/a0019902

Butler, A. C., Karpicke, J. D., & Roediger, H. L. I. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13*(4), 273–281. https://doi.org/10.1037/1076-898X.13.4.273

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1563–1569. https://doi.org/10.1037/a0017021

Carpenter, S. K. (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*(6), 1547–1552. https://doi.org/10.1037/a0024140

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*(2), 268–276. http://dx.doi.org/10.3758/BF03193405

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, *13*(5), 826–830. https://doi.org/10.3758/BF03194004

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36*(2), 438–448. https://doi.org/10.3758/MC.36.2.438

Carpenter, S. K., & Yeung, K. L. (2017). The role of mediator strength in learning from retrieval. *Journal of Memory and Language*, *92*, 128–141. https://doi.org/10.1016/j.jml.2016.06.008

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633–642. https://doi.org/10.3758/BF03202713

Ebbinghaus, H. (1913). *Memory: A Contribution to Experimental Psychology*. New York, NY, US: Teachers College, Columbia University (Reprinted Bristol: Thoemmes Press, 1999).

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*(3), 392–399. https://doi.org/10.1037/0022-0663.81.3.392

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments & Computers*, *16*(2), 96–101. https://doi.org/10.3758/BF03202365

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory*, *19*(3), 290–304. https://doi.org/10.1080/09658211.2011.560121

Howard, M. W., & Kahana, M. J. (2002). A Distributed Representation of Temporal Context. *Journal of Mathematical Psychology*, *46*(3), 269–299. https://doi.org/10.1006/jmps.2001.1388

Jang, Y., & Huber, D. E. (2008). Context retrieval and context change in free recall: Recalling from long-term memory drives list isolation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 112–127. https://doi.org/10.1037/0278-7393.34.1.112

Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: The role of retrievability.

The Quarterly Journal of Experimental Psychology, *65*(5), 962–975. https://doi.org/10.1080/17470218.2011.638079

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-Based Learning: An Episodic Context Account. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 61, pp. 237–284). Academic Press. Retrieved from http://www.sciencedirect.com/science/article/pii/B9780128002834000071

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*(2), 151–162. https://doi.org/10.1016/j.jml.2006.09.004

Keresztes, A., Kaiser, D., Kovács, G., & Racsmány, M. (2014). Testing Promotes Long-Term Learning via Stabilizing Activation Patterns in a Large Network of Brain Areas. *Cerebral Cortex*, *24*(11), 3025–3035. https://doi.org/10.1093/cercor/bht158

Lehman, M., & Karpicke, J. D. (2016). Elaborative Retrieval: Do Semantic Mediators Improve Memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000267

Lehman, M., & Malmberg, K. J. (2013). A buffer model of memory encoding and temporal correlations in retrieval. *Psychological Review*, *120*(1), 155–189. https://doi.org/10.1037/a0030851

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*(6), 1787–1794. https://doi.org/10.1037/xlm0000012

MacLeod, C. M., & Nelson, T. O. (1984). Response latency and response accuracy as measures of memory. *Acta Psychologica*, *57*(3), 215–235. https://doi.org/10.1016/0001-6918(84)90032-5

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*(4–5), 494–513. https://doi.org/10.1080/09541440701326154

Mensink, G.-J., & Raaijmakers, J. G. W. (1988). A model for interference and forgetting. *Psychological Review*, *95*(4), 434–455. https://doi.org/10.1037/0033-295X.95.4.434

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533.

Murdock, B. B., & Okada, R. (1970). Interresponse times in single-trial free recall. *Journal of Experimental Psychology*, *86*(2), 263–267. https://doi.org/10.1037/h0029993

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, *110*(4), 611–646. http://dx.doi.org/10.1037/0033-295X.110.4.611

Pan, S. C., Gopal, A., & Rickard, T. C. (2016). Testing With Feedback Yields Potent, but Piecewise, Learning of History and Biology Facts. *Journal of Educational Psychology*, 563–575. https://doi.org/10.1037/edu0000074

Pan, S. C., Wong, C. M., Potter, Z. E., Mejia, J., & Rickard, T. C. (2015). Does test-enhanced learning transfer for triple associates? *Memory & Cognition*, *44*(1), 24–36. https://doi.org/10.3758/s13421-015-0547-x

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*(1), 3–8. https://doi.org/10.1037/0278-7393.31.1.3

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*(1), 129–156. https://doi.org/10.1037/a0014420

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*(2), 93–134. https://doi.org/10.1037/0033-295X.88.2.93

Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The Power of Successive Relearning: Improving Performance on Course Exams and Long-Term Retention. *Educational Psychology Review*, *25*(4), 523–548. https://doi.org/10.1007/s10648-013-9240-4

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006b). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science*, *17*(3), 249–255. http://dx.doi.org/10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., & Tulving, E. (1979). Exclusion of learned material from recall as a postretrieval operation. *Journal of Verbal Learning & Verbal Behavior*, *18*(5), 601–615. https://doi.org/10.1016/S0022-5371(79)90334-7

Rohrer, D. (1996). On the relative and absolute strength of a memory trace. *Memory & Cognition*, *24*(2), 188–201. https://doi.org/10.3758/BF03200880

Rohrer, D., & Wixted, J. T. (1993). Proactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(5), 1024–1039. http://dx.doi.org/10.1037/0278-7393.19.5.1024

Rohrer, D., & Wixted, J. T. (1994). An analysis of latency and interresponse time in free recall. *Memory & Cognition*, *22*(5), 511–524. http://dx.doi.org/10.3758/BF03198390

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Thomas, R. C., & McDaniel, M. A. (2013). Testing and feedback effects on front-end control over later retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 437–450. https://doi.org/10.1037/a0028886

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology*, *56*(4), 252–257. https://doi.org/10.1027/1618-3169.56.4.252

van den Broek, G. S. E., Segers, E., Takashima, A., & Verhoeven, L. (2014). Do testing effects change over time? Insights from immediate and delayed retrieval speed. *Memory*, *22*(7), 803–812. https://doi.org/10.1080/09658211.2013.831455

van den Broek, G. S. E., Takashima, A., Segers, E., Fernández, G., & Verhoeven, L. (2013). Neural correlates of testing effects in vocabulary learning. *NeuroImage*, *78*, 94–102. https://doi.org/10.1016/j.neuroimage.2013.03.071

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. https://doi.org/10.1016/j.jml.2014.04.004

Watkins, O. C., & Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory, 1*(4), 442–452. https://doi.org/10.1037/0278-7393.1.4.442

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11*(6), 571–580. http://dx.doi.org/10.1080/09658210244000414

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, *20*(1), 6–10. https://doi.org/10.3758/BF03202594

Wixted, J. T., & Rohrer, D. (1994). Analyzing the dynamics of free recall: An integrative review of the empirical literature. *Psychonomic Bulletin & Review*, *1*(1), 89–106. http://dx.doi.org/10.3758/BF03200763