March 2020

Testing the convergent retrieval learning theory of testing effects

William J. Hopper
University of Massachusetts Amherst

Testing the convergent retrieval learning theory of testing effects

A Dissertation Presented

by

WILLIAM HOPPER

Submitted to the Graduate School of the

University of Massachusetts Amherst in partial fulfillment

of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2020

Psychological and Brain Sciences

Testing the convergent retrieval learning theory of testing effects

A Dissertation Presented

by

WILLIAM HOPPER

Approved as to style and content by:

_____

David E. Huber, Chair

_____

Jeffrey Starns, Member

_____

Agnès Lacreuse, Member

_____

Ken Kleinman, Member

_____

Caren Rotello, Department Head
Psychological and Brain Sciences

ABSTRACT

Testing the convergent retrieval learning theory of testing effects

February 2020

WILLIAM J. HOPPER, B.S., UNIVERSITY OF CALIFORNIA SAN DIEGO

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: David E. Huber

What is learned from retrieving a memory that is not learned by studying the same information? In response to this question, I have proposed a new theory of retrieval-based learning in which I argue that retrieval strengthens the ability to completely activate all portions of a memory trace from an initial state of partial activation. In effect, retrieval serves to unitize the features of a memory, making the entire memory remain retrievable in the future when cue-related activation may be weaker. This theory, called the Primary and Convergent Retrieval (PCR) model, explains why practice tests produce both better long-term retention and faster retrieval than restudy of the same information. In this dissertation, I explore and test several predictions arising from the assumptions of the PCR model's learning rule. In Experiment 1, I use evidence from retrieval latencies to demonstrate that even unsuccessful retrieval attempts produce learning. In Experiment 2, I demonstrate retrieval practice does not generalize between retrieval cues, which has important consequences for assumptions about what the features of memory representations may be, and retrieval routes through these features. And in Experiment 3, I show that when the same gradual unfolding of features that is assumed to allow learning during retrieval is deliberately engineered to occur during encoding, it produces the same

types of retention and latency benefits produced by retrieval. These experiments further support the PCR model by confirming its prediction about what is learned from testing, and when this learning may be expected. Portions of this dissertation have appeared in previously published works; specifically, much of the general introduction and general discussion has appeared in Hopper & Huber (2018), and the entirety of Experiment 1 is also reported in Hopper & Huber (2019). Experiments 2 and 3 have not yet appeared in any published volume.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

LEARNING FROM RETRIEVAL AND THE TESTING EFFECT

1.1. Introduction

It is rare that a single exposure to new information, like a new colleague's name, is sufficient for anyone to retain that new information over a long period of time. Rather, experience tells us that retention requires repetition and practice, and even then, forgetting is difficult to stave off. If long-term retention of information is a goal everyone shares (at least from time to time), it is useful to ask whether any method of re-exposure produces better retention than others. Research beginning close to one hundred years ago has provided an answer to this practical question: recalling information from your memory is a highly effective method for retaining that information, and produces better long-term performance than methods involving restudy (see Roediger & Karpicke, 2006a; and Karpicke, 2017 for recent reviews). Results demonstrating the long-term retention benefits of testing are so ubiquitous and consistent that the term "the testing effect" has been coined as shorthand to describe the universal finding of better memory accuracy after a practice test as compared to restudying the same information.

A classic example of the testing effect comes from an experiment by Roediger and Karpicke (2006b), who contrasted the effects of restudy and free recall test practice on memory for facts from textbook passages. When their participants' memory was tested five minutes after learning the material (i.e., a brief retention interval), memory was better for facts that were restudied than for facts that were practiced with a free recall test. However, this pattern of results reversed when the final test was given after a longer retention interval (two days or seven days); memory was better for facts that were

practiced with a free recall test than for facts that were restudied. The exact pattern of results from this experiment is shown in Figure 1, and similar patterns of results have been shown across multiple other studies (Carpenter, Pashler, Wixted, & Vul, 2008; Kuo & Hirshman, 1996; Toppino & Cohen, 2009; Wheeler, Ewers, & Buonanno, 2003). Thus, if you are cramming for an exam just before it starts, the best strategy may be quickly re-reading 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.

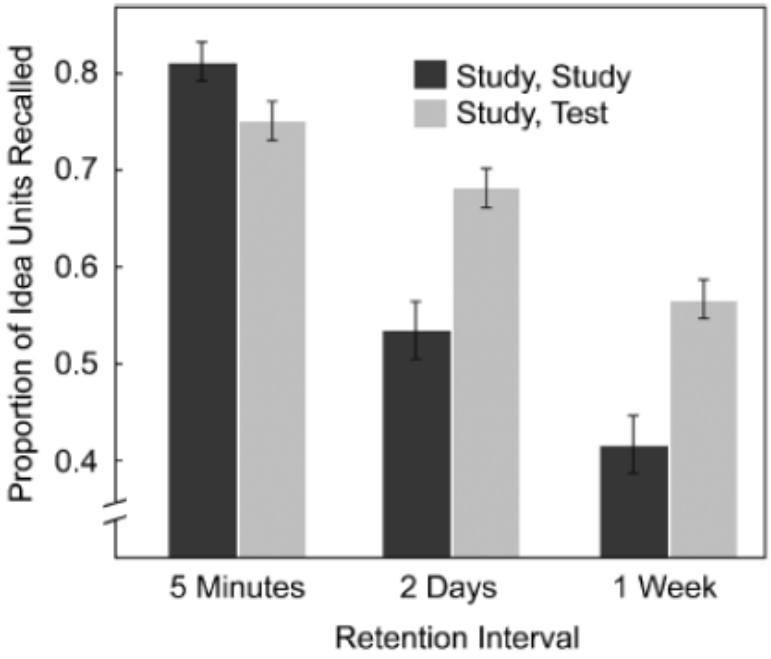

Figure 1: Recall accuracy in Experiment 1 of Roediger and Karpicke, 2006b. Restudying produces better accuracy at short retention intervals, but test practice produces better accuracy at longer retention intervals. Error bars represent standard errors of the mean in each condition.

Findings like this give rise to the general conclusion that restudying and recalling information produce unequal learning about that information. But despite many empirical investigations and several proposed theoretical accounts, there is still no universally accepted explanation for testing effects, or retrieval-based learning more broadly (Rowland, 2014). However, a useful framework for conceptualizing the different effects

of restudy and test practice comes from the "bifurcated" distribution account presented by Kornell, Bjork and Garcia (2011). Under this account, the "early" advantage for restudy and "late" advantage for test practice stems from the effects these practice methods have on the properties of the memory strength distributions for to-be-remembered items. A chance to restudy increases the memory strength for *all* restudied items, effectively shifting the distribution of memory strengths up to a higher mean without changing its shape. A practice test (without feedback) produces a great deal of learning for the items recalled on the practice test, but no learning for the unrecalled items. This divides the memory strength distribution into two portions – the part strengthened by retrieval, and the residual, unstrengthened portion. If the strengthening from restudy is less than the strengthening from successful recall practice, and not all items are recalled on the practice test, this produces better performance in the short term following restudy (because all of the restudied items are strengthened, instead of just the already recallable items) but better performance in the long term following test practice (because the larger increase from successful retrieval makes the items remain memorable for a longer period of time).

The notion of a bifurcated distribution of memory strengths following test practice has been supported by several studies examining its predictions. Jang et al. (2012) administered an initial practice test for all studied items for the purposes of dividing items into separate pools of recallable and non-recallable items. Items in each pool underwent either additional study or test practice, before taking a final test on all items. The results showed that the advantage of restudy before the final test was almost entirely due to strengthening the items in the non-recallable pool. In addition, if recall

success on the practice test is very high (i.e., there is little bifurcation of the memory strength distribution), 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 magnitudes of learning reflect different processes. The theory set out here specifies the learning mechanisms underlying the testing effect by building upon the common assumption in formal models of memory that recall is a two-stage process.

1.2. Two-Stage Retrieval Operations

Many formal process models of memory assume that recall is a two stage process with the following stages: 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. In developing an account of the benefits from taking a recall practice test, learning for different kinds of associations is considered (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 2, 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

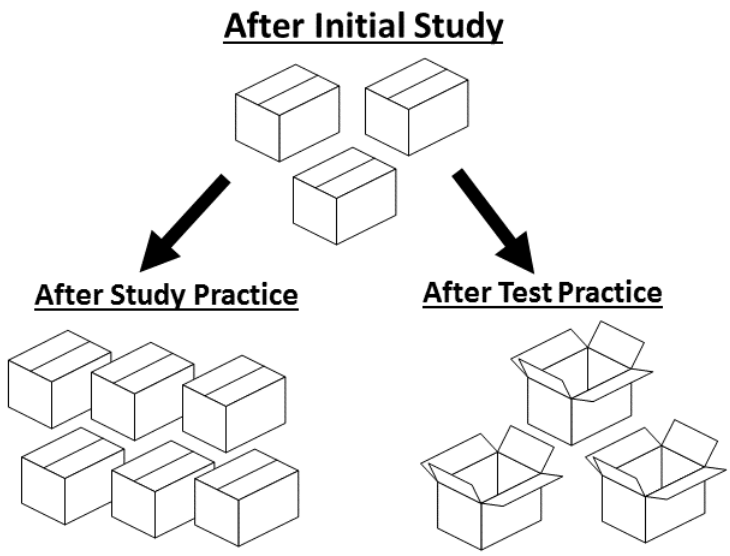

Figure 2: This diagram shows the hypothesized effect of study practice and test practice using a shipping warehouse where packages must be retrieved and opened as an analogy for the memory system. Restudy strengthens the retrieval cues, resulting in a large set of correctly retrieved memory packages (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).

are included in the search set, which increases recall accuracy) whereas recall practice doesn't 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 this account deviates from prior memory models in proposing that the two retrieval operations are differently affected by different kinds of practice, new terminology is used to described these two processes: 'primary retrieval' versus 'convergent retrieval'.

1.3. Primary and Convergent Retrieval

The Primary and Convergent Retrieval (PCR) model of recall makes 3 core assumptions about how information is recalled. The first two relate to the distinct

stages of processing involved in recalling information, while the third the learning rule of the model. The first assumption is that there is an initial stage of recall, termed Primary Retrieval. In this 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). The second assumption is that the initial stage of Primary Retrieval is complemented by a second stage, termed Convergent Retrieval. This subsequent process activates the initially dormant features in one of the items (presumably the most active item). It is considered to be a separate process from Primary Retrieval because features is not activated by retrieval cues directly, but rather by associations between the features within the item (i.e., intra-item associations). 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 R. Brown & McNeill, 1966). However, if this process succeeds, all of the features become active and the item is available for report (i.e., recall is successful).

The third assumption describes when and what is learned during retrieval. The PCR model makes the assumption of 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.

The unique contribution of the PCR model is further specification of the recovery process and a proposal for the types of learning (e.g., recall practice) that uniquely affect recovery. Previously proposed memory models assume that the probability of recovery is based on the same item strength that underlies sampling and familiarity. In contrast, the PCR model allows for associative information that is unique to item recovery. Furthermore, unlike previous memory models, the PCR model assumes that item recovery takes some time, with this duration affected by recall practice.

An example of convergent retrieval is shown in Figure 3A 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 3B).

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 dormant features makes it more likely that attempting recall with the original retrieval cues will be successful. This intra-item learning also serves to reduce retrieval latency because less work is required in convergent retrieval. For example, in Figure 3C, 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).

Why should delay result in activation of a subset of the originally activated features? During initial study, features representing the current context are active before item features, resulting in directional associations from the context to the item's features. As previously outlined, these new associations provide the basis of primary retrieval. An increase in retention interval is thought to weaken the match between the context of initial learning and the context at time of retrieval (Howard & Kahana, 2002; Mensink & Raaijmakers, 1988). This change in context means that some features of the learning context (study and test practice) will not be present at the delayed final test. Without these features being present, fewer of the originally learned target features are activated Furthermore, a change of context cannot spontaneously result in the activation of target features that were not associated with the context prior to the delay.

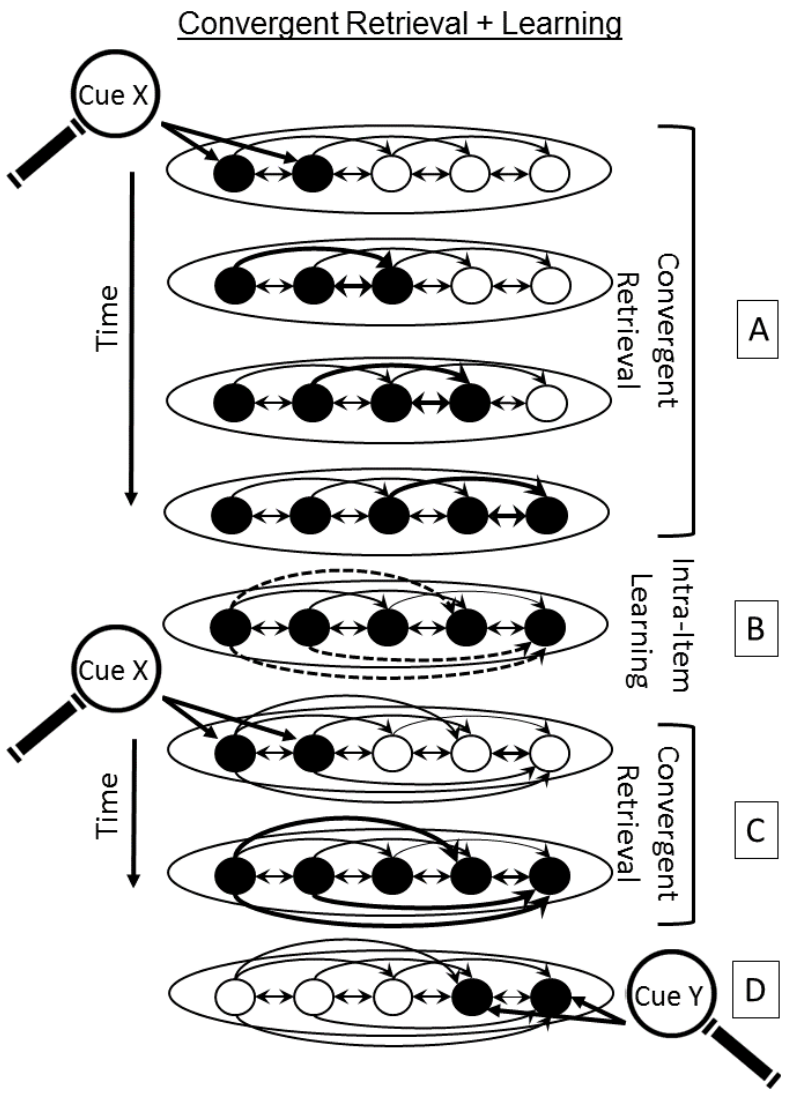

Figure 3: The operations and intra-item learning that occur during convergent retrieval. A) The gradual activation of an item's features during convergent retrieval, initiated by primary retrieval based on retrieval cue X in the first time step. With two lines of support (indicated by the bold arrows), 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.

Thus, the features activated by primary retrieval after a delay (i.e., a context change) are a subset of the originally learned features. Because of this weakened activation from the context cues, intra-item learning is critical for retrieval success -- prior recall practice makes it more likely that convergent retrieval will succeed from this subset of features, thus protecting recall performance from context changes.

The benefits of recall practice can be contrasted with restudy. Restudy produces new associations between the retrieval cues and the item, but because re-presenting the item activates all its features at once, there is no opportunity for intra-item learning. Because restudy enlarges the set of features activated by primary retrieval, this can increase the probability of convergent retrieval success for a brief retention interval (e.g., it more likely that the starting point will support success). However, over a longer retention interval, this enlarged set of features may be diminished due to a change in context cues to the point that convergent retrieval fails. Thus, restudy result in faster forgetting relative to recall practice, because the learning it produces is more susceptible to contextual changes that occur over time. The PCR model also predicts is that test practice should also produce retrieval on the final test that restudy produces, because of its explanatory mechanism of better retention. If enhanced intra-item associations explain the long-term retention advantage of test practice, and enhanced intra-item associations produce faster retrievals, then test practice should also produce faster retrieval on the final test that restudy produces.

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.

1.4. Support for the Convergent Retrieval Learning Hypothesis

Support for the convergent retrieval learning hypothesis can be found across several lines of research. Prior studies have shown that practice tests involving recall produce larger retention benefits compared to practice using recognition tests (Carpenter & DeLosh, 2006; Glover, 1989; Rowland, 2014)[1]. This difference is important because recall tests, but not recognition tests, require the full recovery of an item's details in order to respond. The recall vs. recognition advantage is consistent with

[1] This pattern of results is often interpreted in favor a retrieval effort explanation of testing effects (i.e., that the effort expended retrieving memories produces greater strengthening than restudying the information). However, this is not mutually exclusive with the PCR model, as retrieval effort explanations don't specify the mechanisms underlying learning from retrieval.

the PCR model's convergent retrieval learning mechanism: when testing requires recovery of detailed features from a state of partial activation, intra-item associations are strengthened and enable robust retention. Prior studies have also shown that recall practice results in faster retrievals on subsequent tests, an *a priori* prediction of the PCR model (Keresztes, Kaiser, Kovács, & Racsmány, 2014; Pyc & Rawson, 2009; van den Broek, Segers, Takashima, & Verhoeven, 2014; van den Broek, Takashima, Segers, Fernández, & Verhoeven, 2013; Vaughn & Rawson, 2014).

Hopper and Huber (2018) performed a detailed investigation of the PCR model's prediction that retrieval should be faster on a final test following test practice than on a final test following restudy, testing this prediction across both free recall and cued recall. Their free recall task measured accuracy and inter-retrieval time (IRT) for recall of 15 item word lists immediately following a practice recall test or a restudy. Accuracy and recall latency dissociated between the two methods of practice; restudy produced higher final test accuracy than test practice, but an analysis of the IRTs revealed faster recall following recall practice, especially for the last few items that were recalled. The results from their cued recall experiment conceptually replicated the free recall results, but also included additional conditions testing whether recall practice with one cue would generalize to a final test with a different cue. Each target item was paired with two unique cue words during initial study, but was only practiced again with one of these cues. The cue given on the final test was either the practiced cue or the unpracticed cue, and final test recall was found to be faster only with the practiced cue, regardless of retention interval. This result is compatible with the PCR model because of the directional nature of associative learning in the model; intra-item learning beginning from

a set of features activated by one cue may not benefit retrieval when convergent retrieval proceeds from a *different* set of features activated by another cue (i.e., learned navigation from one start point doesn't necessarily help navigation to the same goal from a different start point). Thus, the results of Hopper and Huber provide support for the PCR model, confirming its central predictions about faster recall latency following test practice.

1.5. Generality of Retrieval-Based Learning

In addition to investigating long-term retention benefits, researchers have also investigated whether the learning from retrieval is transferable. In other words, does the learning from retrieval (whatever it may be) help you answer other, related questions, or are the benefits restricted to better recall of the *specific* answer produced during test practice? Experiments examining transfer of learning have found that tests can produce more general learning than restudy, but generalization benefits are not as universal as the classic testing effect. For example, Butler (2010) showed that retrieval practice of concepts from science-related passages produced better transfer of knowledge to questions from different domains than restudy (e.g., answering questions about the structure and function of animal wings to generalized to answering questions about the structure and function of aircraft wings), and McDaniel et al. (2013) showed that practice recalling the definition of a term generalized to recalling the term given the definition on the final test[2]. However, Hinze and Wiley (2011) and Pan, Gopal and Rickard (2016) found that retrieving one component of a learned fact did not increase the ability to retrieve another component of the fact (e.g., recalling that it was Thomas Jefferson who

[2] These studies also included feedback during the practice test.

purchased the Louisiana Territory from Spain did not increase the ability to recall who Jefferson purchased the territory *from*). Thus, it appears that testing benefits generalization more so when knowledge application is required, and less so when simply needing to recall adjacent information.

Given that the benefits of retrieval practice do not seem completely boundless, examining situations when retrieval practice impacts memory other than the ability to recall the same information across two identical tests may be a useful method for testing the proposed theoretical mechanisms for producing the classic testing effect. The PCR model makes a simple prediction about the general circumstances when memory should be enhanced: when information is gradually activated, and when similar information is activated by the retrieval cues used across tests Here, I report three experiments arising from this prediction. In each, memory is predicted to be enhanced when these conditions are met. These experiments demonstrate that convergent retrieval leaning can be both highly general, applying across different classes of items and encoding methods, but also be highly specific, requiring consistency between learning conditions and testing conditions. In Experiment 1, I demonstrate that similar learning from testing occurs for unretrieved and retrieved items alike. In Experiment 2, I demonstrate that all cued recall practice is *not* equivalent, suggesting that unique conjunctions representing cue-target pairs are part of the features stored in episodic memory traces. And in Experiment 3, I demonstrate that learning from tests generalizes beyond tests – that is, when a study procedure mimics the convergent retrieval process, it produces a retention benefit similar to a practice test. Together, these experiments further support the convergent retrieval learning mechanism and its putative role in promoting retention following testing by

demonstrating the PCR model's power to predict what is learned from testing, and when this learning may be expected.

CHAPTER 2

EXPERIMENT 1

2.1 Learning from the failure to recall

The learning rule assumed in the PCR model concerns the temporal activation of features regardless of whether full convergence is achieved. Thus, the PCR model also makes predictions regarding the effects of recall failure, which are illustrated in Figure 4. In this example, the items on the left and right have different sets of pre-existing intra-item connections. While both items have connections supporting the gradual activation of additional features, the convergent retrieval process ultimately reaches a dead end for the item on the right; the second feature only has one intra-item connection and thus does not become active even if all the other item features are active. This results in a "tip-of-the-tongue" state in which the test-taker may be keenly aware that they know the answer, but is unable to overtly produce the answer owing to a missing feature. Nevertheless, because this dead-end was gradually reached, intra-item learning occurs, as indicated by the dashed arrows. This intra-item learning supports faster failure to recall on the final test, assuming that the test-taker decides to give up on the process at the point when the pattern of features no longer changes. Again by analogy to navigation, when attempting to go from location A to location B, one might encounter an insurmountable road block, and this experience will make it easier/faster to navigate from A to the road block in the future (i.e., faster failure).

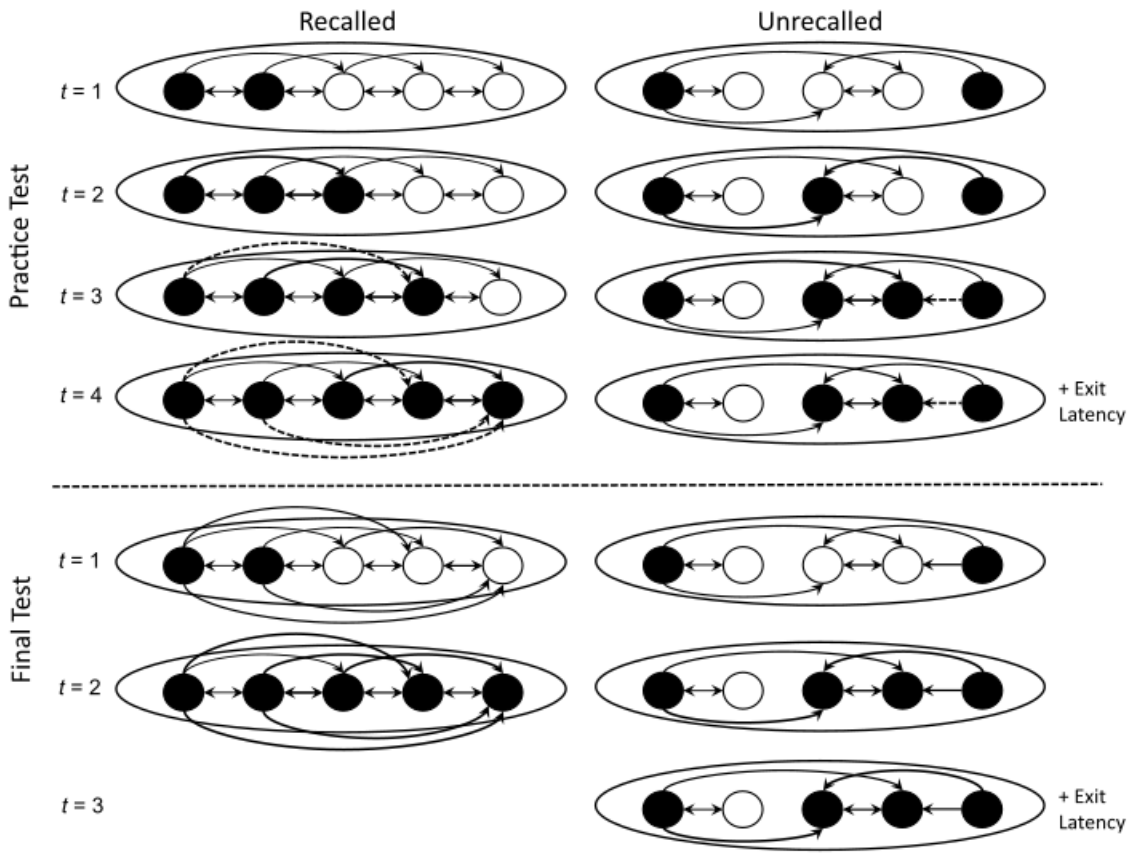

Figure 4: Comparison of the convergent retrieval process and intra-item learning for both recalled and unrecalled items. Left Column: Successful recall on a practice test produces intra-item learning, enabling faster retrieval on the final test. Right Column: Unsuccessful recall on a practice test also entails intra-item learning, but in this case intra-item learning results in faster failure to recall on the final test.

Unfortunately, the large retrieval practice literature does not provide a test of these predictions because failure-to-recall latencies are rarely examined. The primary goal of the current study is collection of these failure latencies to test these predictions. I refer to the key dependent measure as the failure latency rather than error latency because these predictions concern a stalling of the recall process and, subsequently, the decision to give up on the attempt (i.e., errors of omission). This type of error can be contrasted with recalling the wrong item (i.e., errors of intrusion or commission). The PCR model's predictions for intrusions and commissions is complicated, depending on whether the same incorrect answer is given on the practice and final tests (in which case these error

latencies should be faster following recall practice for exactly the same reason that correct recalls are faster) or whether intrusion/commission errors are unique to the final test (in which case they may reflect a failure of primary retrieval). In light of these complexities, intrusion/commission errors are not considered in this experiment. Instead, I focus on recall failures (omissions) by requiring participants to make a binary "recall" or "can't recall" decision on each cued recall trial (see Figure 5A). If the participant reports that they can recall the missing target, they are instructed to immediately type in the target word (i.e., they need to have recalled the word before pressing the "recall" button). Each participant was tested on a large number word pairs in this fashion, across multiple days, allowing analysis of latency distributions.

Thus far, memory models have not addressed learning from cued recall practice as measured with recall latencies, although different aspects of this situation have been investigated in isolation. Raaijmakers and Shiffrin (1981) implemented learning from retrieval in the SAM model by increasing the associative strength between retrieval cues and memory traces following recall success, and this assumption explained part-list cueing effects as well as output interference in free recall. Similarly, Criss, Malmberg, and Shiffrin (2011) modeled output interference during recognition tests with the REM model, assuming that existing memory traces are updated when items are judged to be old, whereas new traces are added to the memory set when items are judged to be new. Most similar to the current experiment, Nobel and Shiffrin (2001; also see Diller, Nobel, & Shiffrin, 2001), modeled intrusion and "give up" latencies from cued recall testing of word pairs using a modified version of the REM model. However, Nobel and Shiffrin did not test the same word pairs more than once, and so these data are unsuitable for testing

the prediction that a failure to recall following one cued recall test should produce a faster failure to recall on a subsequent cued recall test with the same cue word. In brief, existing memory models have not addressed learning from recall failures. This is not to say these models are incompatible with learning from errors of omission, or that they make different predictions than the PCR model, but rather that their predictions in this respect have not been explored.

2.2 Recall latencies: a change in response bias or memory?

The PCR model predicts that the convergent retrieval process should unfold more quickly after recall practice and, furthermore, that this speed-up should exist both for recall success and recall failure. However, an examination of average recall latency could be misleading in regard to these predictions. More specifically, faster responses can occur at the expense of accuracy (i.e., a speed-accuracy tradeoff) in a situation where there is a shift in response bias rather than a change in the evidence accumulation process. In the context of cued recall, such a response bias corresponds to the adoption of a more liberal stopping rule (e.g., giving up more readily) or a more conservative stopping rule (i.e., careful checking before producing an answer), with changes in this stopping rule depending on the cue item. For instance, the participant may explicitly remember that they failed or succeeded with the cue word on the practice test and use this knowledge of their own memory process to adjust the effort they are willing to expend during cued recall. Such behavior is a kind of 'metamemory' (i.e., knowledge about your memory), similar to 'judgments of learning' (T. O. Nelson & Dunlosky, 1991), if the setting of response bias is based on a prediction for the outcome of the ongoing recall attempt. Similarly, the familiarity of the cue word may influence the time it takes to make a

"Recall" vs. "Can't Recall" decision. For example, Malmberg (2008) showed that increasing the familiarity of a cue can increase the amount of time spent searching memory without increasing accuracy. Fortunately, cue familiarity is not expected to differ between the restudy and test practice conditions, given that both re-present the cue word.[3] Nonetheless, an adjustment of response bias based on memory for the success of the previous recall attempt is likely to affect performance and thus it is critical that any analysis of the results consider this decision making aspect of the task.

This decision making component lies outside of the current scope of the PCR model. In situations such as this, measurement models are often used to untangle decisional aspects (e.g., response bias) from memorial aspects (e.g., sensitivity) of the data, such as with an application of signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 2005). Fortunately, a broad class of reaction time measurement models (so-called 'sequential sampling' models) have been developed to address the potential ambiguity of a speed-accuracy tradeoff and these models have proven useful in the study of memory (S. D. Brown & Heathcote, 2008; Ratcliff & Smith, 2004; Starns, 2014). These decision models address latency and choice data on a trial-by-trial basis to identify whether an on-average change in speed or accuracy reflects a change in response bias versus a change in the evidence accumulation process. Following in this tradition, a sequential sampling model is used to transform the data into more psychological relevant parameters that, for instance, indicate whether the observed latency effects reflect the

[3] If test practice increases cue familiarity more than restudy, this would work *against* the central prediction of faster "Can't Recall" decisions after test practice, assuming cue familiarity increases willingness to continue searching memory.

retrieval process (as predicted by PCR) or whether they reflect a change in the speed-accuracy tradeoff (or more likely some combination of these factors).

2.3 The Linear Ballistic Accumulator Model

In a sequential sampling model, "evidence" builds up in support of the possible response options over the course of the decision process (Donkin, Brown, & Heathcote, 2011). These models have been applied to recognition memory tasks, assuming that the drift rate parameter reflects access to information from the memory system (Osth, Bora, Dennis, & Heathcote, 2017; Ratcliff, 1978; Ratcliff & Starns, 2009; Ratcliff, Thapar, & McKoon, 2004). As applied to recall, I assume that the drift rate reflects activation of item features during the convergent retrieval process – each additional feature that is activated provides further evidence towards a response.

There are many successful sequential sampling models (see Voss, Nagler, & Lerche, 2013 for a good introduction to the properties of these models), all of which include parameters that capture the speed-accuracy tradeoff in which participants can elect to respond slowly and accurately, or quickly but with more errors. More specifically, some parameters of these models are related to response bias (e.g., the required evidence threshold and/or the starting level of evidence), which affect latency distributions in a different manner than parameters related to the rate of information accrual (i.e., drift rate). In the current case, if the test-taker realizes that a cue was previously used in a practice test, they may adopt a lower threshold for making an educated guess of the target or a lower threshold for giving up on the recall attempt. In addition, changes to primary retrieval may provide a higher level of initial evidence. These effects can be contrasted with intra-item learning, which should change the rate of

information accrual over the course of the retrieval attempt. A test of these predictions requires separate measurement of parameters related to recall success versus parameters related to recall failure, and so I used the Linear Ballistic Accumulator (LBA) model (S. D. Brown & Heathcote, 2008), because it is an independent race model, with one racer capturing recall success and the other capturing recall failure.

The adoption of the LBA model was made after careful consideration. For example, a diffusion model could have applied to these data (Ratcliff, 1978; Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002). A diffusion model uses a single accumulator (and thus, a single drift rate parameter), and response outcomes are determined by which of two opposing decisional boundaries is reached first by the accumulator. In the Ratcliff diffusion model, the two responses are not independent of one another; any evidence gained in favor of one response is evidence against the other. Thus, the parameters of the diffusion model would not allow identification of characteristics of recall failure separate from the characteristics of recall success. Because the LBA model assumes independent accumulators, it is possible to identify parameters unique to the recall failure process as well as the recall success process. It is important to note that while evidence accumulators in the LBA model are independent, I am not asserting that there are two independent convergent retrieval processes occurring simultaneously. Rather, the accumulators of the LBA are being used to *measure* the internal evidence supporting recall of the target item and the evidence supporting the failure to recall.

The LBA model assumes that any decision between a set of alternatives is based on the outcome of a race between competing evidence accumulation processes. Each accumulator begins with some initial amount of evidence supporting the corresponding

response alternative. Over time, additional evidence is gained, until one accumulator reaches a critical threshold, and at that point in time, the corresponding response is given. Thus, the LBA model describes the decision process as a race between evidence accumulators towards a response threshold, as illustrated in Figure 5B. The intercept of the vertical axis and the bold line shows the amount of initial evidence, the upward slope of the bold line shows the accumulation of evidence over time, and the dashed horizontal line across the top represents the critical amount evidence that must be reached for that particular response alternative.

Because the LBA is an independent race model, the amount of evidence for the "Recall and "Can't Recall" alternatives may have two different initial values, increase at two different rates, and be racing towards two different thresholds for responding. The initial evidence value for each alternative is drawn from a uniform distribution on each trial, ranging from 0 to the parameter A (thus beginning at a value of $A/2$ on average). The rate of evidence accumulation follows a normal distribution across trials, with mean v and standard deviation s. The v parameter is commonly referred to as the drift rate, as it describes the average speed at which evidence strength "drifts" away from the initial starting value over the course of a trial. The amount of evidence required for a specific response (i.e., the response threshold parameter b) is assumed to be constant across trials, as is the amount of time required for non-decisional processes necessary to give a response (e.g., planning and executing motor movements), which is given by the parameter T_0. The model is described as being "linear" because it assumes that evidence is accumulated at a constant rate and this accumulation is "ballistic" in the sense that once the starting points and drift rates are determined, the accumulation process has a

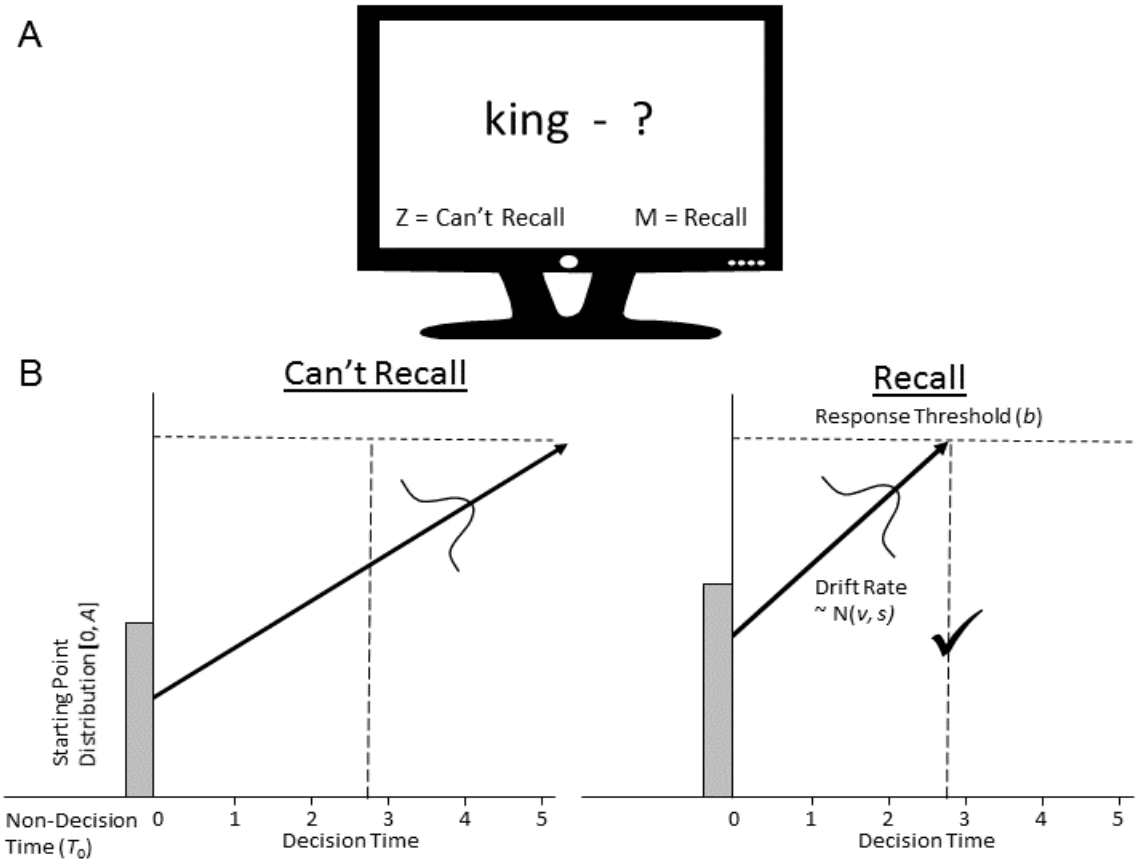

Figure 5: Panel A: Example of the "Recall" vs "Can't Recall" decision made on each test trial of the current experiment. Panel B: Schematic representation of a decision between the "Recall" and "Can't Recall" alternatives as described by the LBA model. The accumulator that intersects the response threshold first is the chosen alternative, and the response time is the amount of time elapsed before the response threshold is met. In this example, the "Recall" accumulator reaches the threshold first, and is the response given on this simulated trial.

deterministic conclusion. These characteristics can be contrasted with the assumption of random-walk fluctuations of evidence within a trial, made by models such as the drift-diffusion model. However, the assumption of a linear and ballistic evidence accumulation process is made primarily for reasons of mathematical convenience and simulation studies have demonstrated that the LBA's parameters are largely similar to those of the drift-diffusion model (Donkin, Brown, Heathcote, & Wagenmakers, 2011).

2.4 The LBA as applied to recall latencies

In a typical application of the LBA model, the accumulators represent a small set of different possible response options: E.g., is the stimulus a word or a non-word (S. D. Brown & Heathcote, 2008), or is the stimulus moving to the right or the left (Forstmann et al., 2008). In contrast, the set of possible words to recall is vast. My assumption in using the LBA model to test the predictions of the PCR model is that the test-taker gives a "Recall' response once they have recalled a particular word, but gives a "Can't Recall" response once the convergent retrieval process stalls with no further change in the pattern of item feature activations. Thus, the goal of applying the LBA is to describe the nature of recall success versus recall failure generically, rather than describing recall of a particular word. With this goal in mind, the drift rate of the "Recall' accumulator is predicted to increase as a function of prior successful recall practice (i.e., the benefits of successful test practice) whereas the drift rate of the "Can't Recall" accumulator is predicted to increase as a function of prior recall failure (i.e., the cost of unsuccessful test practice). In addition, I use the LBA to test alternative accounts of the data, examining, for instance, whether the results are better explained by changes in the response thresholds versus changes in the drift rates (see Figure 5B).

Trial to trial variability in the starting point of the evidence accumulation process is necessary to capture the fast errors that occur when speed is emphasized over accuracy (S. D. Brown & Heathcote, 2008). The parameter governing this starting point distribution, *A*, can also be utilized to represent bias towards a specific response alternative; an accumulator with starting points sampled from distributions with a larger *A* parameter will also have a higher mean starting point and thus start closer to the

response boundary on average. However, the *A* parameter is not typically used in this fashion, and most applications of the LBA model set the *A* parameter to the same value for all accumulators. The current situation is different because the choice behavior being modeled is not a simple stimulus classification. In the PCR model, the starting point of the convergent retrieval process is the set of features activated by the retrieval cues. If primary retrieval is very rapid, taking essentially the same duration on each trial, the starting point for the evidence accumulation process as measured by the LBA model might reflect the primary retrieval strength in response to the retrieval cues. Under this assumption, different values of the *A* parameter for different accumulators and experimental conditions are possible; according to the PCR model, primary retrieval should differ across conditions, corresponding to different starting points for convergent retrieval. Alternatively, if primary retrieval is itself a dynamic process, then primary retrieval and convergent retrieval will collectively serve to specify the drift rate. In this case, an LBA model that attempts to capture the data through variation in the starting point parameter will fail.

Faster responding owing to a change in response bias can be distinguished from faster responding owing to a change in the retrieval process because each possibility produces a different change in the shape of the latency distribution. For instance, if drift rate for the recall accumulator increases, this will increase accuracy and decreases mean latency, and the nature of this decrease is a less variable, more normally distributed recall latency distribution. In other words, the convergence process happens more quickly, with greater reliability. If the recall threshold decreases, this also produces an increase in accuracy and decrease in mean latency. However, in this case, the recall latency

distribution becomes more exponentially shaped rather than normally shaped. For instance, with a sufficiently low recall threshold, on some trials the test-taker decides immediately to respond "Recall" (e.g., primary retrieval places the evidence above the threshold level), but failing this, it is still possible that a positive, but near-zero drift rate will take a long time to reach the threshold. In practice however, these differences are likely to be subtle, and so model comparison is used to determine which parameter changes provide the best account of the data.

To characterize the nature of any changes in retrieval latency and accuracy, I fit several different LBA models, allowing only the *A* (starting point), *v* (drift rate) or *b* (decision boundary/threshold) parameter to vary between conditions, which instantiate different hypotheses about the effects of restudy and test practice. If the benefits of a practice recall test increase primary retrieval, this should increase the average evidence accumulation starting point, and the *A* parameter LBA model should provide the best account of the data. I also considered an alternative version in which primary retrieval affected the starting time of evidence accumulation (the T_0 parameter), to examine the possibility that primary retrieval includes not only a strength component, but a latency component (the time before the onset of convergent retrieval) that may vary across items with different primary retrieval strengths. For instance, if primary retrieval activates relatively many features, it may be that that this step occurs more quickly as compared to a situation with weaker primary retrieval. If the benefits of a practice recall test occur because the participant is biased to quickly respond "recall" owing to a change in response bias, this should correspond to a decrease in the evidence threshold, and the *b* parameter LBA model should provide the best account of the data. Finally, as predicted

by the PCR model, if the benefits of a practice test increase the average rate of evidence accumulation during convergent retrieval, this should correspond to an increase in the drift rate, and the *v* parameter LBA model should provide the best account of the data. Furthermore, this should be true not only for faster recall after success on the practice test but also faster failure after failure on the practice test. Finally, several hybrid models were applied that allowed combinations of these parameters to vary across the experimental conditions.

2.5 Summary of Experiment 1

Faster correct recall following recall practice has been observed in several prior studies (Hopper & Huber, 2018; van den Broek et al., 2014), but to date, the effect of recall practice on the speed of recall failure has not been examined. The current study addresses this by examining the latency of recall/can't recall judgments, replicating the finding that recall success results in faster recall success on a subsequent test and also testing the novel prediction of the PCR model that recall failure on a practice test results in faster recall failure on a subsequent final test. By including many trials per participant in each condition, a reaction time measurement model is applied, characterizing the nature of any on-average changes in latency to determine whether these effects reflect a change in drift rate rather than response bias, as expected if the latency change reflects a more rapid convergent retrieval process, rather than a meta-memory strategy to require less evidence to choose the 'recall' response.

2.6 Materials

2.6.1 Participants

Ten individuals were recruited from the University of Massachusetts Amherst community via electronic mailing lists and word of mouth. Participants were compensated at a rate of $15 per hour, plus a $5 bonus for showing up to each session. All participants completed all sessions, and were paid a total of $70. The relatively small sample size reflects an emphasis on collecting enough observations from each participant for participant-level model fitting.

2.6.2 Materials

Twelve hundred (1200) English words were used, with the constraint that each word had between four and 10 letters, and a word frequency between five and 200 uses per million words according to the SUBTLX$_{US}$ corpus (Brysbaert & New, 2009). Words of a single conjugation were selected, and nouns were permitted to be either singular or plural with only one form or the other included in the stimulus set. From this pool, 600 randomly determined word pairs were created, but these same word pairs were used for all participants. These word pairs were grouped into 25 lists of 24 pairs and the pairs within each lists were randomly assigned to the conditions for each subject, with the constraint that four pairs from each list were assigned to each of the six conditions.

2.6.3 Procedure

The experiment was administered over the course of four sessions on four consecutive days. During the first session, participants learned and practiced the first nine of the 24 word pair lists. One third of the pairs in each list were restudied, one third of the

pairs were given a practice cued recall test, and the remaining third were not practiced again after the initial study opportunity. Half of the words in each practice condition were given a final cued recall test immediately after the learning and practice phase for that list, with a brief distractor task (30 seconds of cumulative addition problems) interposed between the practice phase and the final test. The remaining half of the items were given a final cued recall test at the start of Session 2. For both immediate and delayed final tests, the order of items within each list was randomly shuffled. The delayed test was a test of multiple lists, and the test list was blocked by list, with the order of the tested lists the same as the order in which they were studied the previous day.

Following the final test for items from Session 1, participants learned and practiced word pairs from eight new lists during Session 2. Just as in Session 1, word pairs were evenly divided between the three practice conditions (cued recall, restudy, and no practice), half of the word pairs in each condition received a final test immediately, and the remaining half were tested at the start of Session 3 the following day. Following the final test for items from Session 2, participants learned and practiced word pairs from the final eight lists during Session 3.

Again, word pairs learned in Session 3 were evenly divided between the three practice conditions, half of the word pairs received a final test immediately, and the remaining half were tested during Session 4 the next day. No new pairs were learned during Session 4, thus the delayed final test concluded the experiment. A diagram outlining the schedule of learning and testing over the course of the four sessions is shown in Figure 6. This procedure yields a two by three within-subjects factorial design, fully crossing retention interval (immediate vs. delayed final test) with practice type

(cued recall, restudy, and no practice), with 100 final test trials per participant in each condition.

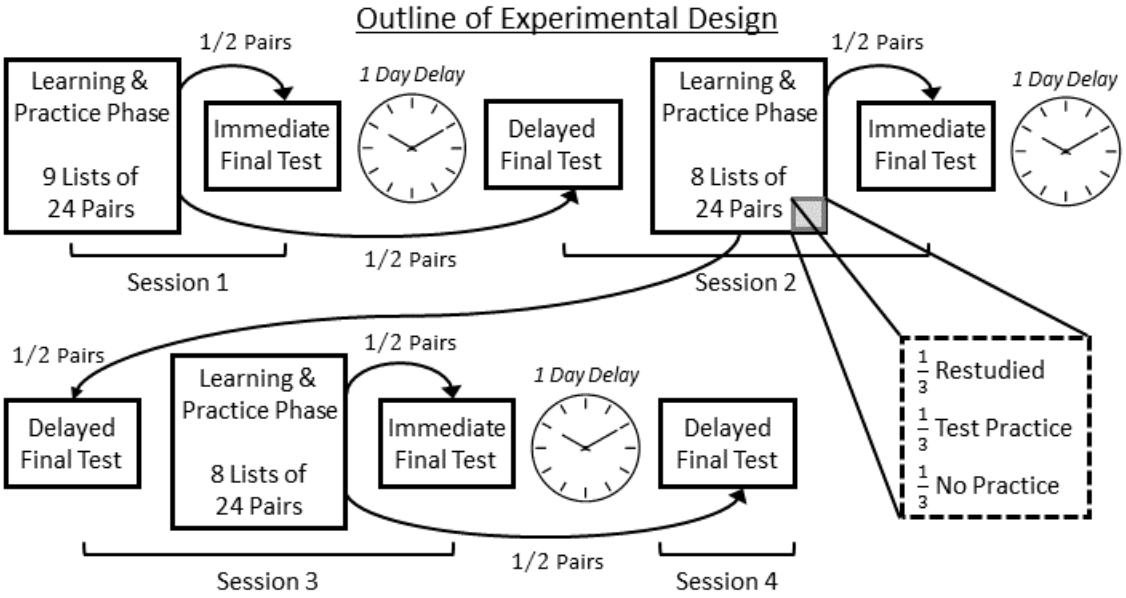

Figure 6: Outline of the experimental design, with the temporal structure of the sessions flowing left to right, and then top to bottom.

During study and restudy trials, word pairs were presented on the computer screen for four seconds, with one word (the "cue" word) on the left, and the other word (the "target") on the right. Test trials (practice and final) had two phases. First, participants were presented with the cue word alone, and had to report whether they could recall the missing target word, or whether they could not recall the missing word. Participants reported this decision by pressing keys on the keyboard (pressing the M key for "Remember" and the Z key for "Don't Remember"). Participants were instructed to only press the "Remember" key when they absolutely knew what the missing word was and were ready to begin typing it in. If participants responded in the affirmative, they were asked to type in the correct word using the keyboard, and press the Enter key to confirm their response. There was a half second inter-stimulus interval between all types of trials (study, restudy and test trials).

To maximize recall performance on the practice tests, the initial study phase and practice phase were intermixed. After every three new word pairs studied, the word pair studied four trials ago was practiced (i.e., restudied or given a cued recall test). Items assigned to the "No Practice" condition were skipped. If necessary, filler word pairs (i.e., pairs that were never practiced or given a final test) were inserted at the end of the list to maintain the lag-3 spacing between "true" pairs in the list. The maximum number of filler pairs that were inserted in any list was three.

2.7 Results

2.7.1 Scoring

All latencies were measured as the duration between the onset of the cue word, and the key press indicating the recall/can't recall decision. A trial was deemed correct only if the participant indicated they could recall the missing target word and subsequently typed the correct word. The accuracy of subsequent typed responses on trials where participants indicated they could recall the target were scored by a software routine that allowed for small misspellings (e.g., letter transposition, pluralization) to be labeled as correct.

2.7.2 Statistical Analysis

The percent of words correctly recalled in each condition is shown in the top panel of Figure 7. Differences in accuracy between conditions were assessed with a logistic mixed-effects regression model, using the *lme4* (Bates, Mächler, Bolker, & Walker, 2015) and *afex* (Singmann, Bolker, Westfall, & Aust, 2018) packages for the R statistical computing environment (R Core Team, 2017). Practice type (restudy, cued

recall, and no practice) and retention interval (immediate vs. delayed final test), as well as their interaction, were included as fixed effects. The model also included random intercepts and slopes for participants in each condition, and random intercepts for each item. This random effects structure was reached by starting with the maximal random effects structure (i.e., random intercepts and slopes for both participants and items in each condition), removing terms from the random effects structure (beginning with the item component) until the model fitting routine was able to converge on stable parameter estimates (see Barr, Levy, Scheepers, & Tily, 2013).

The significance of the fixed effects in the model were assessed using likelihood ratio tests[4]. These tests indicated that the full model including both main effects of practice type and retention interval along with their interaction fit the data significantly better than the restricted model without a main effect of practice type ($\chi^2(2) = 17.86$, $p < .001$), better than the restricted model without a main effect of practice type ($\chi^2(1) = 30.73$, $p < .001$), and better than the restricted model without a practice type by retention interval interaction ($\chi^2(2) = 10.92$, $p = .004$). The conclusion drawn from these model comparison tests is that there were significant differences in recall accuracy between the levels of each condition, as well as an interaction between the practice type and retention interval factors.

From inspection of the top panel of Figure 7, it is clear that the main effect of retention interval reflects lower recall accuracy on the delayed final test. Differences in

[4] All models compared using the likelihood ratio test were fit using the maximum likelihood method. The likelihood ratio test is known to be too liberal when the number of participants is low (Luke, 2017), but given that the accuracy effects reported here are regularly observed, the conclusions from this particular test are unlikely to be Type I errors.

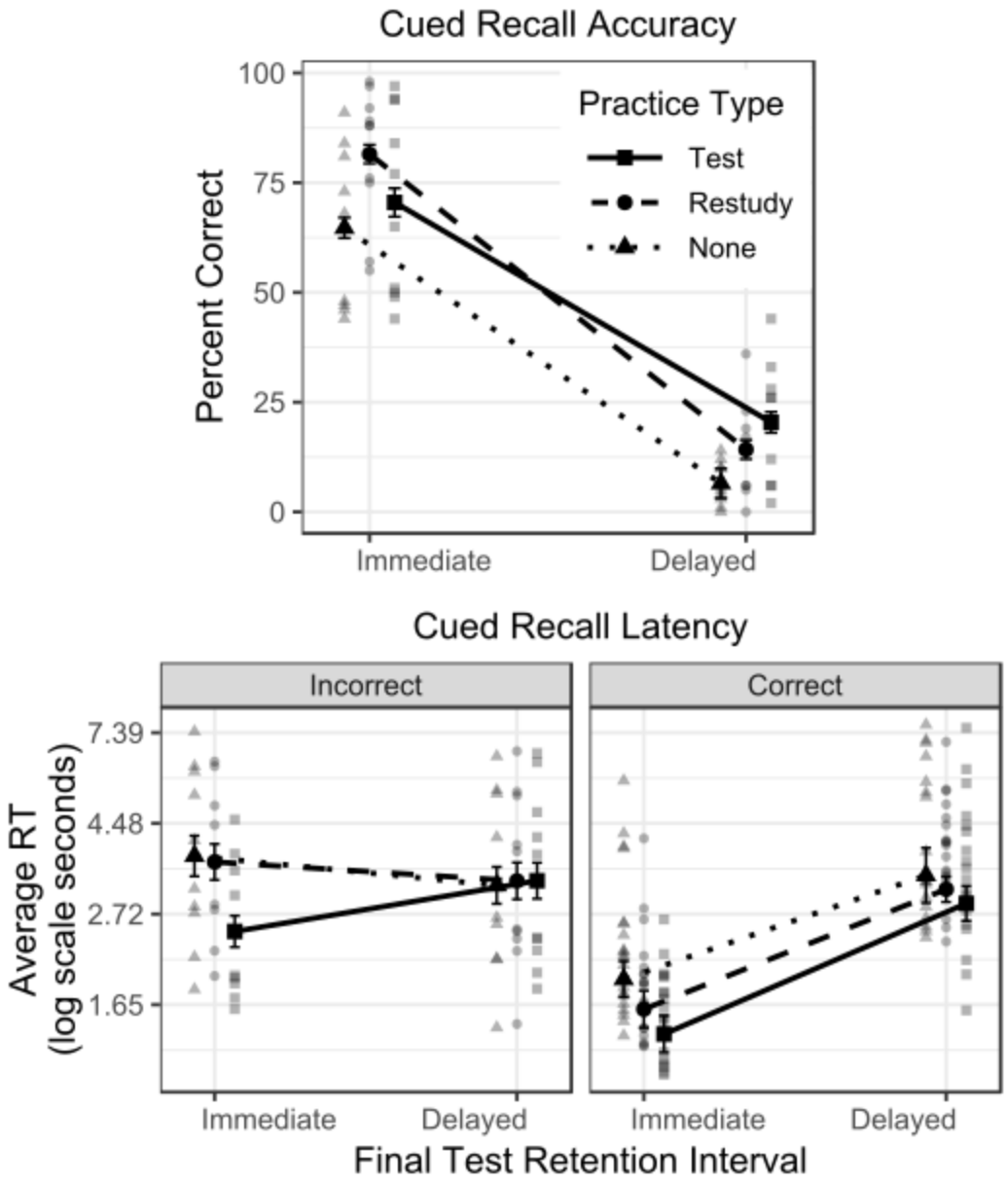

Figure 7: Performance across conditions. Larger, darker points represent averages across participants. Smaller grey points represent observations from individual participants. Error bars represent +/- one standard error of the mean, estimated using the subject-normalized method of Morey (2008). Top Row: Recall accuracy on the final cued recall test. Bottom Row: Average decision latency on the final cued recall test. Incorrect latencies reflect trials where participants indicated they could not recall the target item. Correct latencies reflect trials where participants could recall the target item, and subsequently provided the correct word as a response.

accuracy between practice types were assessed using Holm-Bonferroni corrected contrasts at each retention interval. In the immediate final test condition, accuracy in the no practice condition was significantly below both the restudy condition (z = -6.89, p < .001) and the test practice condition (z = -2.56, p = .01), while performance in the restudy condition was significantly higher than in the test practice condition (z = 3.61, p < .001). In the delayed final test condition, the relationship between the restudy and test practice conditions reversed, with the test practice condition displaying significantly higher accuracy than restudy (z = 2.37, p < .017). Accuracy in the no practice condition was still significantly below both the restudy condition (z = -3.52, p < .001) and the test practice condition (z = -6.06, p < .001) at the delayed final test.

The decision latency for recall and can't recall judgements is shown in the bottom panel of Figure 7. Trials where participants indicated they could recall the missing target word, and subsequently typed in the correct word were considered to be correct recall latencies, with latency determined by the time to press the "remember" key. Only trials where participants indicated they could not recall the missing target word in the decisional phase were treated as error latencies. Trials where participants indicated they could recall the target word, but failed to type in the correct word (7% of final test trials), and trials where recall decision latencies were over 10 seconds (5.8% of trials) were not analyzed. Together, these criteria excluded 11% of the data.

Differences in recall decision latencies between conditions were assessed with a linear mixed-effects regression model, again using the *lme4* and *afex* packages for the R statistical computing environment. Recall decision latencies were log transformed prior to analyses to meet the assumption of Gaussian residual variance in the regression model.

For ease of interpretability, correct and incorrect latencies were analyzed with separate regression models. Practice type and retention interval, as well as their interaction, were included as fixed effects in both models. Both models included random intercepts and slopes for participants in each condition, and both models were fit using the residual maximum likelihood (REML) method.

The significance of the fixed effect components in each model were evaluated with an ANOVA using the Kenward-Roger approximation of the error degrees of freedom in all *F*-tests and follow up contrasts (Kenward & Roger, 1997). For the correct recall decision latencies, there was a significant main effect of practice, $F(2, 6.46) = 8.24$, $p = .016$), a significant main effect of retention interval $F(1, 7.77) = 185.38$, $p < .001$, but no interaction between retention interval and practice type, $F(2, 5.89) = 0.74$, $p = .51$). From the bottom right panel of Figure 7, it is clear that the main effect of retention interval reflects slower responding on the delayed final test for all practice types. Differences in correct recall decision latencies between practice type conditions were assessed using Holm-Bonferroni adjusted contrasts, collapsing over the retention interval factor. Correct responses in the no practice condition were significantly slower than in the test practice condition, $t(7.6) = 4.21$, $p = .009$, and slower than in the restudy condition, $t(7.14) = 2.13$, $p = .069$, though the difference narrowly missed the threshold of statistical significance at $\alpha = .05$. Correct responses in the test practice condition were faster than in the restudy condition, $t(7.61) = 2.72$, $p = .054$), again narrowly missing the threshold of statistical significance.

For the incorrect trial decision latencies (i.e., “Can’t recall” responses), there was no main effect of retention interval, $F(1, 8.25) = 0.04$, $p = .84$, though there was a

significant main effect of practice type, $F(2, 4.73) = 6.89$, $p = .039$, and a significant practice type by retention interval interaction, $F(2, 4.56) = 8.10$, $p = .031$). The nature of the interaction was investigated using Holm-Bonferroni adjusted contrasts between the practice types at each retention interval. There was no difference between the incorrect trial decision latencies for the no practice and restudy condition on the immediate final test, $t(3.87) = .205$, $p = .84$. However, incorrect trial decision latencies in the test practice condition were significantly faster than in both the no practice condition, $t(6.37) = -4.56$, $p = .009$, and the restudy condition $t(4.6) = 3.74$, $p = .031$. There was no difference in incorrect trial decision latencies between any of practice type conditions on the delayed final tests (all $|t|$ statistics < 1). Thus, the main effect and interaction observed in the *F* tests were driven by significantly faster "Cant' Recall" responses in the test practice condition on the immediate final test.

2.7.3 LBA Model Analysis.

Six LBA models with different parameter constraints were applied to the recall decision latencies from the 10 participants individually. Four of these models assessed whether just one of the key parameters could capture the differences between conditions: 1) Convergent retrieval, corresponding to the *v* parameter; 2) Primary retrieval, corresponding to the *A* parameter; 3) an alternative formulation of primary retrieval in which primary retrieval affected the starting time of evidence accumulation, corresponding to the T_0 parameter; and 4) A metamemory change in the response threshold, corresponding to the *b* parameter. A fifth model allowing both primary retrieval and convergent retrieval was considered (both *v* and *A*), to examine whether primary retrieval might load onto the *A* parameter when the *V* parameter also varied. The

models allowing only A and to T_0 vary between conditions performed poorly (were never the winners in model comparison), and the v and A model faired almost as poorly (was preferred only for a few subjects, and only when using the AIC measure of goodness of fit). Within the framework of the PCR model, this suggests that primary retrieval is a dynamic process, similar to convergent retrieval, in which case the drift rate parameter reflects the combined actions of primary and convergent retrieval. The sixth model allowed drift rate and boundary (v and b) to vary across conditions, in order to determine whether the predicted drift rate parameter value differences would still hold when allowing boundary to change as well. To preview the conclusion from this exercise, differences between practice conditions were always explained best by models with free drift rates, and the need for different response boundaries across conditions varied from participant to participant.

All models set the drift rate variance parameter to a constant value (s = .5 for each accumulator). Unless otherwise specified, all models used a common parameter value across all conditions and all accumulators for each free parameter of interest (e.g., for most models, the same value for T_0 was assumed for the recall and can't recall accumulators in all conditions). Test trials for items recalled on the practice tests were modeled separately from test trials for items not recalled on the practice test. Thus, different parameters were allowed for these two types of items, effectively treating them as observations from separate conditions. This follows directly from the assumption of a bifurcated distribution in which the learning processes following successful recall practice are different than the learning from the failure to recall.

The "*v* Free" model allowed separate drift rate parameters (v_0 and v_1) for each accumulator in every condition (i.e., each combination of retention interval and practice type). Separate starting point parameters (A_0 and A_1) were fit for each accumulator, but these parameters were shared across all conditions. The boundary parameter for both accumulators was fixed to a constant value ($b = 4$) across all conditions. Note that this parameterization of the LBA model solves the scaling problem[5] by fixing the response boundary, rather than setting the sum of drift rates to a fixed constant, as is common in the response time modeling literature (Donkin, Brown, & Heathcote, 2009). This model included 19 free parameters per participant (four practice conditions – no practice, restudy, correct test practice and incorrect test practice – times two retention intervals times two accumulators equals 16 drift rate parameters, plus two starting point parameters and the non-decision time parameter).

The "*A* Free" model allowed separate starting point parameters (A_0 and A_1) for each accumulator in every condition. Separate drift rate parameters (v_0 and v_1) were used for each accumulator at each retention interval, but these parameters were shared across all practice type conditions at each retention interval. The boundary parameter for both accumulators was fixed to a constant value ($b = 4$) across all conditions. In total, the "*A* Free" model allowed for 21 free parameters per participant.

The "*b* Free" model allowed separate boundary parameters (b_0 and b_1) for each accumulator in every condition. Just as with the "*A* Free" model, separate drift rate

[5] The parameters of the LBA and Diffusion models can be multiplied by a constant without changing the RT distributions. Fixing a single parameter makes the model identifiable and enables model comparison (analogous to coding a categorical predictor with K levels in a regression model using K-1 dummy coded variables).

parameters were used for each accumulator at each retention interval, but were shared across practice type conditions. The starting point distribution parameter for both accumulators was fixed at a constant value (A = 1.5) across all conditions. This model also allowed 21 free parameters per participant.

The "T_0 Free" model allowed separate non-decision time parameters for all racers and conditions. Like the "A Free" and "b Free" models, separate drift rate parameters were used for each accumulator at each retention interval, but were shared across practice type conditions. Similarly, each racer was allowed a free starting point parameter which was shared across retention interval and practice type conditions. The boundary parameter for both accumulators was fixed to a constant value (b = 4) across all conditions. In total, this model also 22 free parameters per participant.

The "v and A Free" model, allowed separate drift rate parameters (v_0 and v_1) and starting point variability parameters (A_0 and A_1) for each accumulator in every condition. The boundary parameter for both accumulators was fixed to a constant value (b = 4) across all conditions. Similar to the "v and A Free" model, the "v and b Free" model allowed different boundary parameters (b_0 and b_1) for each accumulator across all conditions and different drift rate parameters (v_0 and v_1) for each accumulator across all conditions. Just as in the "b Free" model, the starting point parameter A was fixed at a constant value across all conditions (A = 1.5). Both of these models allowed 33 free parameters per participant.

The actual number of free parameters used was lower than the maximum number possible for some participants because of variation in individual performance levels. For example, some participants failed to recall any items from the no practice condition on

the delayed final test, obviating the need for a drift rate parameter for fitting correct recall latencies from that condition. More generally, models were not fit to conditions where there were fewer than two observations per recall decision alternative.

All models were fit to the data from individual participants separately, using the maximum likelihood method to estimate the best fitting model parameters. The *rtdists* package (Singmann, Brown, Gretton, & Heathcote, 2017) for the R statistical computing environment was used to compute the LBA model's density, quantile, and cumulative distribution functions. Prior to estimating each model's parameters, excessively long decision latencies (> 10 seconds) were removed, and trials with a decision latency in the most extreme 2.5% of the latency distribution (in both tails) were removed for each subject. These exclusion criteria resulted in the elimination of 9% of the trials. Model parameters that maximized the likelihood function were estimated using a box-constrained gradient descent search algorithm. The goodness of fit for each of the six models was compared using both BIC (Bayesian Information Criterion; Schwarz, 1978) and AIC (Akaike Information Criterion; Akaike, 1974). The AIC and BIC are likelihood based statistics that impose a penalty on a model's likelihood proportional to the number of free parameters to account for the flexibility afforded by each free parameter. The AIC and BIC differ in the degree of penalty applied, with AIC typically imposing a smaller penalty than the BIC, thus leading the AIC to favor more complex models than the BIC.

The AIC and BIC for the best fitting parameters of each models are shown in Table 1.

Using the AIC penalty for flexibility, the "*v* Free" was the favored model for participants 2 and 5, and the "*v* and *A* Free" model was favored for participant 10. The "*v*

and *b* Free" model was favored for participants 1, 3, 4, 6, 7, 8, and 9. When considering a simultaneous fit of all participants, the "*v* and *b* Free" model had the lowest total AIC of all six models (**Σ** AIC = 17856.72). The BIC penalty for flexibility tended to favor the single parameter models and the "*b* Free" model was preferred for participants 3, 8 and 9, while the "*v* Free" model was preferred for participants 1, 2, 4, 5, 7, and 10. When considering a simultaneous fit all participants, the "v Free" model provided the lowest total BIC over all subjects (**Σ** BIC = 18923.44).

The difference in the complexity penalty imposed by the AIC and BIC measures, and thus their discrepancy when applied to these models, stems from the goal of each measure. The AIC assesses generalization (i.e., a frequentists test, predicting future data), whereas BIC is based on model selection (i.e., a Bayesian test of hypotheses based on the extant data). In the current case, if you sought the simplest explanation of the data (i.e., BIC), the behavior of most participants was best explained by changes in drift rate. Furthermore, across the entire dataset a change in drift rate was the clear winner under the BIC measure. However, if your goal was to predict future performance (i.e., AIC), using all possible mechanisms, including ones that captured a lower proportion of the variance in the data, using the model with freedom in both the drift and response boundaries would be the best choice for most participants. This model is also the clear winner when considering total AIC across the entire dataset.

In summary, as predicted by the PCR model, a change in drift rate (convergent retrieval) is the most crucial aspect of these results, although there is evidence that response boundaries (response bias, such as with a change in metamemory) changed as well. Crucially, these are not mutually exclusive explanations and while the PCR model

predicted changes in convergent retrieval, it did not specify whether response bias might or might not change as well.

Beyond quantitative assessment of the models, I examined the qualitative pattern of model fits and differences in best-fitting parameter values across experimental conditions to assess the behavior of the three best-fitting models (the "*b* Free", "*v* Free" and "*v* and *b* Free" models).

The "*b* Free" model was favored for participants 3, 8, and 9 under BIC and examination of these participants revealed that they had nearly perfect accuracy on the immediate final test (between 90% and 100% correct) and thus very small differences in accuracy between experimental conditions. However, this model qualitatively misfit the data from the other seven participants, with the best-fitting parameters producing better accuracy on the immediate final test in the no practice condition than the test practice condition. More specifically, to accommodate faster recall success after success on the practice test, the boundary for the "recall" accumulator was set lower, but this served to reduce accuracy (i.e., a negative testing effect not seen in the data). In summary, it appears that this model can only accommodate the data when practice test accuracy is nearly perfect and it incorrectly produces an accuracy deficit following test practice, making this an undesirable explanation of the results.

A more plausible alternative is that different practice conditions produced different response biases as well as different rates of evidence accumulation. Corresponding to this alternative, "*v* and *b* Free" model was favored for seven of the ten subjects under the AIC statistic. Inspecting the fits of this model to data from individual participants showed that allowing free drift rate parameters in addition to the free

boundary parameters corrected the qualitative misfit of the "*b* Free" model, correctly producing an accuracy increase after test practice, as compared to the no practice condition. To understand which free parameters were important in capturing the data, I performed multiple contrasts on the best-fitting parameter values using paired samples *t*-tests[6]. Remarkably, the boundary parameter did not systematically differ between any pairs of practice conditions at either retention interval, for either accumulator. In general, all *t* statistics for comparisons amongst the boundary parameters were less than 1.87. Thus, while the *b* parameter may have been important for producing high quality fits of the data in general (as determined by AIC), it did not differ in a systematic manner across the conditions of interest. As such, it does not appear that response bias provides an adequate explanation of the highly reliable on-average retrieval latency effects seen in these data.

Finally, consider the "*v* Free" model. The correlation between observed accuracy values and model accuracy values was very high for this model ($r = .995$), and the model captured the absolute accuracy rates observed for each practice type and retention interval, including the crossover interaction between the restudy and test practice condition across the two retention intervals. The specific parameter values for each subject under this model are reported in Table 2 and Table 3, respectively. In brief, this model captured all of the important accuracy trends in the data.

To assess whether the "*v* Free" model captured the important retrieval latency effects, Figure 8 displays the joint quantile-probability plots for this model, showing the

[6] Identical *t*-tests were also performed for the "*v* Free" model, and described in detail in the "Parameter Contrasts" section.

correspondence between the observed and predicted quantiles of the recall decision latency distributions in each condition. The joint quantile values were determined by estimating the response times associated with the .1, .3, .5, .7, and .9 quantiles of the conditional latency distribution for each response (i.e., conditioned on recalled or not-recalled). So-called "defective" distributions (i.e., distributions that accumulate to the observed or predicted level of accuracy) were then produced by weighting these conditional quantiles by the predicted or observed accuracy level, depending whether the distribution being plotted was the model or the observed data. For example, assume the .7 quantile of the "Recall" response conditional latency distribution was predicted to be 2.5 seconds by the LBA model, and the total probability of responding "Recall" was predicted to be .6. Then, the joint quantile corresponding to a "Recall" response with a latency of 2.5 seconds would be .6 × .7 = .42. The joint quantiles values were calculated for each participant individually and averaged together to create the values plotted in Figure 8.

In general, the model captured the shapes of the quantile functions in each condition reasonably well. One notable misfit is the long tail in the model's behavior for the correct recall decision on the delayed final test. The model produced a longer tailed distribution of "Recall" responses as compared to the "Can't Recall" responses, while the reverse is true in the observed data. However*,* the model was not fit to these quantile distribution*s* and was instead fit to the raw trial-by-trial data (the quantiles are examined only as a way of assessing model behavior). If the model had been fit to the quantile distributions, it is likely that it would have produced a shorter tail in this situation, as dictated by the observed quantiles. More to the point, the delayed final test condition had

the most skewed distribution of correct recall latencies, with participants usually taking less than 4 seconds to respond, but occasionally taking around 10 to give a correct response. In a maximum likelihood fit of the trial-by-trial data, these outliers play a huge role, imposing a substantial penalty for parameter values that fail to place probability mass that covers these outliers.

In contrast, a quantile function does not differentiate between a situation in which the slowest 10% of trials occur between 4 and 5 seconds versus one in which the penultimate 9% occurs between 4 and 5 seconds, with the last 1% at 9 seconds. Second, and perhaps more importantly, this qualitative misfit of the data represented a tiny fraction of the data (this was the condition with the worst accuracy, and so there is very little data to indicate the shape of the recall success latency distribution in this condition). Thus, the observed quantile function in this case is highly unreliable. Given that the model captures the study/test cross over interaction, and matches the recall decision latency distributions reasonably well, I am satisfied that the "*v* Free" model describes the data accurately enough to interpret its parameter values.

2.7.4 Parameter Contrasts.

Because the model with different drift rate parameters in each condition was the most consistent winner, I conclude that an adequate description of the data requires different drift rates. As described earlier, the PCR model predicted that the drift rate for the correct ("Recall") accumulator should be highest after successful test practice, and, furthermore that the drift rate for the incorrect ("Can't Recall") accumulator should be

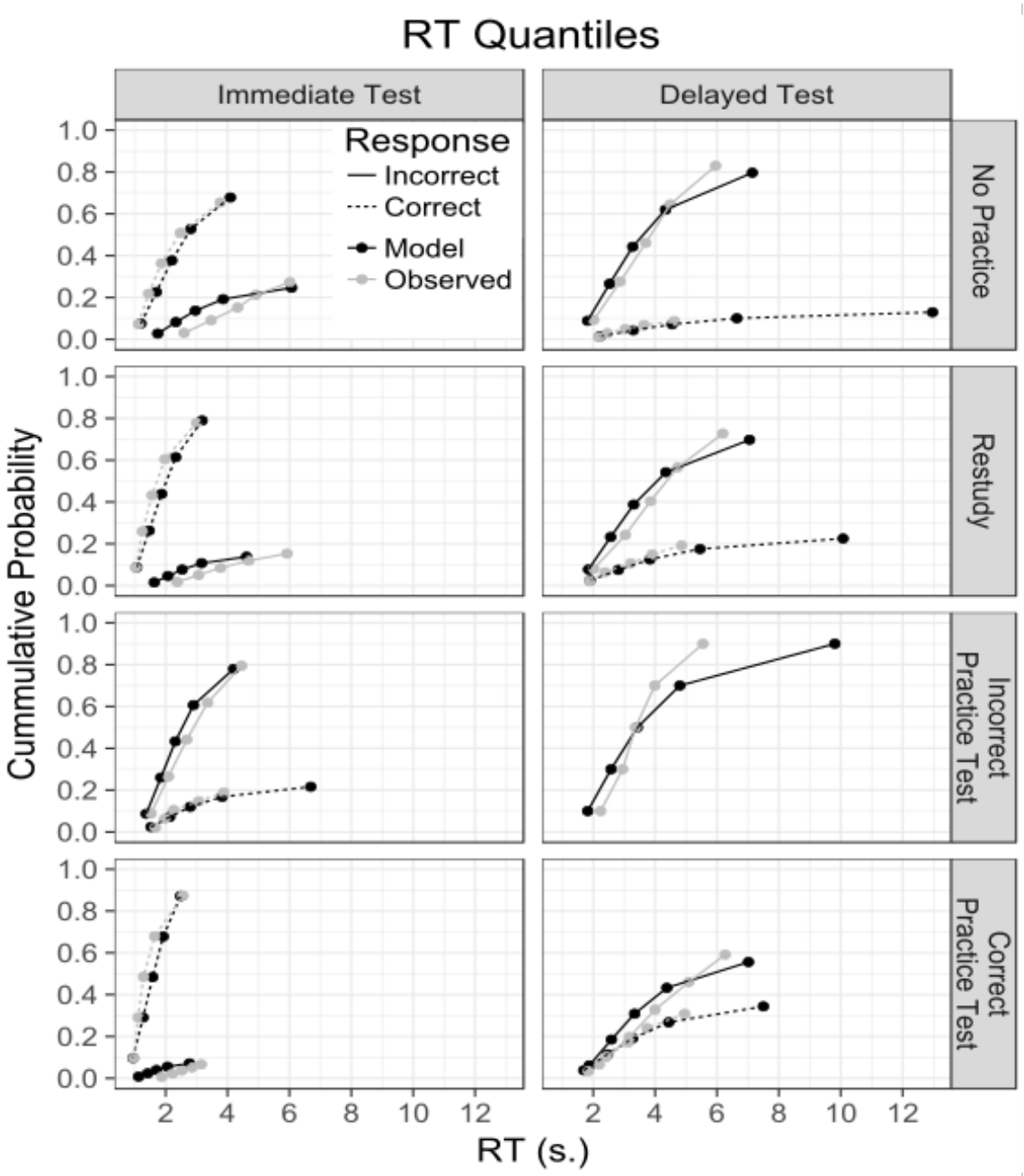

Figure 8: LBA Model RT quantiles, together with empirical quantiles estimated directly from the observed data. The quantile values were estimated at the .1, .3, .5, .7, and .9 quantiles of the RT distributions. No quantile functions for correct "Recall" responses are presented for the delayed final test incorrect practice test condition because there were an insufficient number of responses for this situation.

highest after unsuccessful test practice. To test these predictions I compared the average drift rate parameters in the "*v* Free" model between all pairs of practice types at each

retention interval condition using paired-samples two-sided *t*-tests (the mean values are plotted along the bottom row of Figure 9 for the "*v* Free" model). All *p*-values were corrected for multiple comparisons using the Holm-Bonferroni procedure, treating the contrasts for each drift rate parameter (v_0 and v_1) as separate families of tests[7].

First, consider the "Can't Recall" accumulator's drift rate (v_0), shown in the lower-left panel of Figure 9. Confirming the key predictions of the PCR model, the drift rate on the immediate test following incorrect test practice was greater than for the no practice condition, $t(8) = 7.73$, $p < .001$, and greater than for the restudy condition, $t(9) = 6.09$, $p = .001$. None of the remaining comparisons reached statistical significance (minimum *p*-value = .17). No significant differences were found between any pairs of the "Can't Recall" accumulator's drift rate parameters for different practice types at the delayed final test. The correct test practice condition for the incorrect racer is not shown in Figure 9 considering that there were far fewer trials of this type and also to simplify the figure; a full report of these parameter values appears in Appendix A1. A larger drift rate parameter for "Can't Recall" responses following a previous retrieval failure (i.e., an immediate test of the incorrect test practice condition) supports the PCR model's prediction that the convergent retrieval process will stall more quickly following a prior

[7] Some of the contrasts reported here have identical *p* values because of the mechanisms of the Holm-Bonferroni FWE rate correction. The sequential Holm-Bonferroni procedure tests hypothesis ordered by *p* values, from smallest to larger. The smaller *p* values tested first are subjected to more conservative correction than later tests. This means that one *p* value that was initially slightly smaller than another may become larger after both have been corrected, because of the difference in the correction factor applied to them. In order to prevent the rank order of the *p*-values from being distorted by the correction, the two *p*-values are both corrected to the larger of the two.

failure to recall.

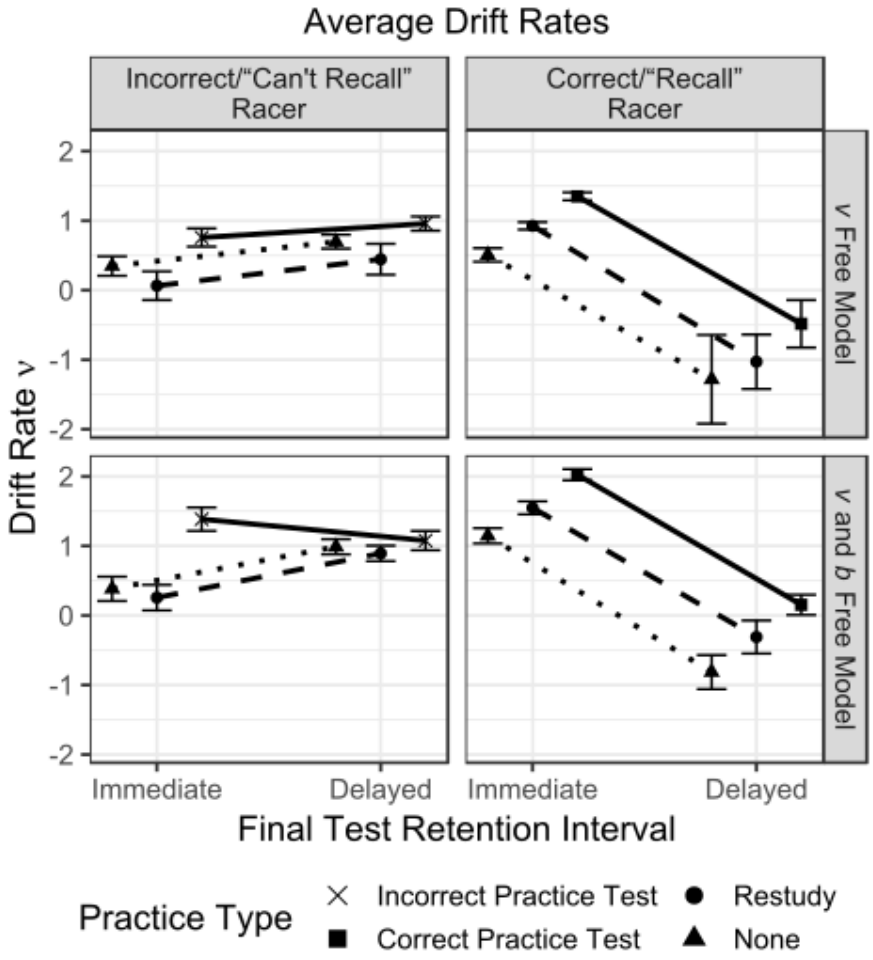

Figure 9: Average drift rates across participants for each condition from the "v Free" (bottom row) and "v and b Free" models (top row). Error bars represent +/- one standard error of the mean within each condition.

Next, consider contrasts involving drift rates for the "Recall" accumulator on the final test, shown in the lower-right panel of Figure 9. All pairwise comparisons between the "Recall" accumulator drift rates (v_1) for the different practice types on the immediate final test were statistically significant. The drift rate for the correct test practice condition was greater than the drift rate for the no practice condition, $t(9) = 6.15$, $p < .001$, and greater than the drift rate for the restudy condition $t(9) = 4.42$, $p = .01$. In turn, the drift rate for the restudy condition was greater than that of the no practice condition $t(9) =$

6.77, $p < .001$, and greater than the incorrect test practice condition, $t(4) = 4.38$, $p = .024$. The drift for the incorrect test practice condition was the smallest of all, significantly less than that of the no practice condition $t(4) = 3.63$, $p = .024$, and the correct test practice conditions $t(4) = 4.65$, $p = .024$. The incorrect test practice condition is not shown in the figure considering that very few trials involved incorrect test practice followed by a correct final test; furthermore, contrasts involving the incorrect test practice condition have fewer degrees of freedom than the other contrasts because not all subjects produced correct responses on the immediate final test for pairs they failed to recall on the practice test. A similar pattern of contrasts was observed for the delayed final test conditions. For the delayed conditions, pairwise differences were found between the "Recall" accumulator drift rate parameters (v_1) for the no practice, restudy and correct test practice conditions. The correct test practice drift rate was significantly greater than the drift rate for the restudy condition, $t(8) = 2.67$, $p = .024$, and significantly greater than the no practice condition, $t(7) = 5.85$, $p = .024$. The restudy condition drift rate was also significantly greater than the no practice condition drift rate, $t(7) = 3.38$, $p = .024$.

In summary, both the mean latency results and the LBA modeling results support the predictions made by the PCR model that successful recall practice results in faster recall success on a final test, while, at the same time, recall failure during recall practice results in faster recall failure on a final test. Furthermore, these effects appear to reflect the retrieval process itself (drift rate) rather than a change in response bias.

2.8 Discussion

The PCR model learning rule specifies that directional associations are learned between already active features and subsequently activated features. This supports the

learning of associations between retrieval cues and the item in the case of initial study and restudy practice. In addition, this same process supports the learning of associations between the features representing an item, but only when item features become active in a gradual fashion, such as occurs during recall practice. Thus, the PCR model provides a novel mechanism for explaining the learning benefits of taking a practice test as compared to passive restudy. In support of this account, the current study confirmed that the faster recall latency observed following a practice test reflects the recall process itself rather than a metamemory response bias to hastily give a "Recall" response based on knowledge that the prior recall attempt was successful.

However, the learning rule of the PCR model does not require complete retrieval success as a condition for strengthening intra-item associations. Thus, learning may take place even if the convergent retrieval process stalls without reaching full convergence (i.e., learning from the failure to recall). This is not learning in the colloquial sense of acquiring new information, but rather learning in the sense of changes in behavior that arise as the result of experience; in this case, what is learned is how *not* to recall the correct target item. If intra-item learning takes place even when retrieval fails, then subsequent retrieval attempts for the same item should reflect this learning in the same way as when retrieval is successful: faster recall (failure) latencies should be observed. The currently study confirmed this prediction both in terms of average failure latencies but also in terms of a change in the recall process itself rather than a metamemory response bias to hastily give "Can't Recall" responses based on knowledge that the prior recall attempt failed.

Error latencies have proven useful for constraining theories of recognition memory (e.g., Cox & Shiffrin, 2017; Starns, 2014), but error latencies are rarely considered in the study of recall (although see Diller et al., 2001; and Nobel & Shiffrin, 2001). One reason for this is that a typical recall task does not ask participants to indicate *when* they've failed to recall. Instead, most recall experiments give participants a fixed recall period, with the failure to recall indicated by the conclusion of this time period without recall. Instead, the current experiment asked participants to report whether they could, or could not, recall the missing target item when given the cue word. These responses are roughly similar to Judgments of Learning (T. O. Nelson & Dunlosky, 1991), although in this case the judgment is immediately followed by typing in the answer if the 'Recall option is chosen. The recall responses from this procedure replicated previous findings of faster recall following test practice from studies that used more traditional measures of recall latency (Hopper & Huber, 2018; van den Broek et al., 2014). This correspondence suggests that participants performed the recall/can't recall decision task in a similar fashion to a standard cued recall paradigm (that is, by recalling the target word before making a response). In addition, this technique confirmed the novel prediction of the PCR model that retrieval failures (i.e., the "Can't Recall" decisions) would also be faster following recall failure on the practice test. A reaction time model of decision making (the LBA model) was applied to these data to determine whether test practice caused a change in the speed-accuracy tradeoff (i.e., a change in response bias) or whether it changed the retrieval process (i.e., drift rate). Comparisons between different LBA models identified that a change in drift rate provided the best

account of the data and, furthermore, the drift rate parameters reliably changed in the predicted manner.

2.8.1 Caveats and Concerns

The PCR model makes no prediction as to whether subjects will or will not adopt a metamemory strategy that adjusts decision thresholds for the recall/can't recall decision. However, regardless of whether such a strategy is adopted, the PCR model predicts that drift rates will change. One subject was best fit by the "*v* and *b* Free" model as assessed by BIC and several others were fit best fit by this model as assessed by AIC, raising the possibility that there were biases and as well drift rate changes. As seen in the top two panels in Figure 9, the pattern of drift rates is nearly identical for this model as compared to the "*v* Free" model (comparing the top graphs to the corresponding bottom graphs). Thus, the drift rate results remain even when allowing for changes in the response boundary. A keen eye might note that the magnitude of the drift rate increase for the immediate final test "Can't Recall" accumulator after failure on the practice test is reduced for the "*v* and *b* Free" model. This suggests that part of the on-average speed-up after recall failure on the practice test is indeed a metamemory decision strategy (although the drift rates still reliably changed). It is important to note however, that the "*v* and *b* Free" model may be too flexible as general model, as indicated by the higher total BIC measure, and by the finding that the threshold parameters did not reliable differ between conditions for this model.

Another concern comes from consideration of item selection effects. The LBA model analyses separated the test practice condition into two pseudo-conditions based on success versus failure on the practice test. However, because this was a post-hoc

separation of the data, this may have introduced item selection effects (i.e., some items are more recallable than others in general, and these pseudo-conditions would select for easy versus hard items). I addressed this concern by directly comparing response latencies on the practice test to response latencies on the final test for two groups of items: items that were not recalled on both tests (i.e., failure - failure items) and items that were recalled on both tests (i.e., success-success items). Providing clear evidence against an item selection effect account of the Experiment 1 results, the average correct recall latency decreased by 1.22 seconds for a successful immediate final test as compared to the same items on a successful practice test ($t(9) = 9.35$, $p < .001$) and the average failure recall latency decreased by 3.93 seconds for recall failure on the immediate final test as compared to recall failure latency for the same items on the practice test ($t(8) = 7.17$, $p < .001$)[8]. The results for the delayed final test are more complicated considering that considerable forgetting occurred over the course of 24 hours, which is likely to make recall success slow (e.g., if asked to recall what you did the summer before last, you probably could, but it would take a while to remember) and at the same time make recall failure fast (e.g., if asked to recall your first birthday, you might immediately state that you can't recall). Thus it is not surprising that the average correct recall latency increased by 1.15 seconds from practice test success to a delayed final test success ($t(9) = 3.76$, $p = .005$), and decreased by 2.14 seconds from practice test failure to a delayed final test failure ($t(8) = 5.27$, $p < .001$).

To further investigate item selection effects, I applied the LBA model to the joint practice and final test data. As with the mean latency analysis, this was done for the

[8] Paired *t* tests were performed on the log scale, to satisfy the assumption of normality.

correct-correct items and the failure-failure items, and so in this case the LBA was used only to describe the shape of the latency distribution, as determined by drift rate, and changes in drift rate from practice to final test, rather than also explaining accuracy (which was by definition perfect or zero for these two groups of items). The model included a unique starting point parameter A for each response type (success or failure) that was shared across test types, and a non-decision time parameter T_0 that was shared across all response and test types. As with the mean latency results, the drift rate for the correct recall latency distribution increased by .85 from the practice test to the immediate final test ($t(9) = 10.91$, $p < .001$), while the drift rate for the correct recall latency distribution decreased by .22 from the practice test to the delayed final test ($t(9) = 2.84$, $p = .019$). The drift rate for the recall failure latency distribution increased by .75 from the practice test to the immediate final test ($t(8) = 18.99$, $p < .001$) and increased by .42 from the practice test to the delayed final test ($t(8) = 3.95$, $p = .008$). Thus, even if the pseudo-conditions selected for different kinds of items, it still appears that recall success on the practice test led to faster recall success on an immediate final test whereas recall failure on the practice test led to faster recall failure on the immediate final test.

The effect of a failed retrieval on subsequent performance has been examined at least once before in the context of the testing effect. Kornell, Klein, & Rawson (2015) had participants study weakly associated word pairs in preparation for two cued recall tests (i.e., practice and then a final test). The key comparison in their study was between two kinds of practice, both of which involved an initial presentation of the cue alone for a recall attempt of the target. In one condition, this initial recall attempt was followed by copying down the correct answer, regardless of recall success (i.e., this served as

feedback), while in the other condition, subjects were given a relatively easy fragment completion of the target after the initial recall attempt such that they could find the answer through their own retrieval processes rather than overt feedback. These conditions produced approximately equal final test performance, which was considerably better than other conditions that involved copying or fragment completion *without* first attempting recall based on the cue alone. However, study through fragment completion was better than copying in the absence an initial recall attempt. This pattern of results indicates that the practice test retrieval *attempt* is the key to effective learning, and that whether the correct answer is reached by recall success (fragment completion) or feedback (copying the target word) is inconsequential. In other words, there is a beneficial effect of recall failure if feedback is provided.

These results are readily explained by the PCR model. According to the PCR model, intra-item learning requires an initial partial activation of the item followed by complete activation of the item, although this complete activation could be achieved either through convergent retrieval or through feedback. However, if there is no initial partial activation (i.e., if there is no initial retrieval attempt) then there is no intra-item learning. Thus, there is beneficial learning from recall failure if that failure is immediately followed with some form of feedback for the item. According to the PCR model, this partial activation might be for the word form itself (e.g., recalling that the first two letter of the correct answer) or it might be for the episodic conjunction created during study (e.g., recalling the mental image created in response to a word pair). In either case, if this partial activation is followed either by either full retrieval, or by feedback for the

correct answer, this will strengthen a directed pathway from the retrieval cues to the answer via this partial activation.

2.8.2 Sequential Sampling Model and Recall Decisions.

Experiment 1 includes a novel application of sequential sampling models, which are more typically used for decision making. Rather than a binary decision about a presented stimulus, participants judged the outcome of the retrieval process – whether it had identified the target word, or whether it had failed to recover the target. In this application of the LBA model, the mapping between the model parameters and psychological variables was slightly different than the typical choice situation. Specifically, the *v* parameter (drift rate) was interpreted as a measure of convergent retrieval strength (and possibly primary retrieval strength) whereas the *b* parameter (threshold) was interpreted in terms of the metamemory process dictating how readily to give up on the retrieval attempt (i.e., the stopping rule). At first glance, this application seems far afield from traditional decision making applications of these models. In light of this, it is worth revisiting the foundational concepts that led to the development of sequential sampling models.

Ratcliff (1978) describes the theory and application of a diffusion model as applied to recognition memory. In doing so, he described the drift rate of the diffusion process (i.e., evidence accumulation) as reflecting the relatedness of the recognition probe to the contents of memory. This relatedness value was conceptualized as the outcome of a feature matching process between the probe and the contents of memory. Specifically, Ratcliff wrote "… probe and memory-set item features are matched one by one. A count is kept of the combined sum of the number of feature matches and non-

matches, so that for a feature match, a counter is incremented, and for a feature non-match, the counter is decremented. The counter begins at some starting value Z, and if a total of A counts are reached, the probe is declared to match the memory-set item …" (Ratcliff, 1978, p. 63). This interpretation of the evidence accumulation process is remarkably similar to the current application of the LBA model as a way of describing the accumulation of item features during the recovery process.

Straying from the original Ratcliff diffusion model, this application of the LBA model also concerned the accumulation of evidence toward the decision to cease the recall attempt, with this potentially exhibiting different dynamics. For recall successes, the decision threshold is reached when the target item is recovered into awareness. For recall failures, several possibilities exist. In the domain of free recall, the decision to cease retrieval attempts is well described by a stopping rule based on the accumulated number of retrieval failures (Dougherty, Harbison, & Davelaar, 2014). Dougherty et al. hypothesized that each retrieval failure involved a new sample from the sample space of potential memories. However, in the case of cued recall, the sample space may play less of a role, such as indicated by the failure to find list-strength effects with cued recall even though such list-strength effects with spaced repetitions are found with free recall (Malmberg & Shiffrin, 2005; Wilson & Criss, 2017). Instead of accumulated failures in the sampling process, the accumulated failures that drive the decision to cease a cued recall attempt may be occurring within the recovery process. In other words, a participant may attempt to "read out" an item from the set of currently active features, and reach their "Can't Recall" decision after some number of failures to name the pattern of features (which is likely to occur if the pattern is no longer converging). Alternatively, a

participant may be directly monitoring changes in the set of activated features, ceasing the retrieval attempt when this set has stabilized without convergence. The results of Experiment 1 are compatible with either explanation, and further experiments are necessary to understand the stopping rule utilized in cued recall tasks.

CHAPTER 3

EXPERIMENT 2

3.1 Semantic Cue Switching

In Experiment 2, I explore the conditions under which generalization between retrieval cues does or does not occur following test practice. In other words: When, if ever, does practice recalling information given one prompt enhance the ability to recall that same information given a different prompt? Recently, Hopper and Huber (2018) investigated transfer between retrieval cues in a cued recall paradigm. In this experiment, participants learned to associate pairs of unrelated words; one word in each pair was designated the cue word and the other word was designated the target word, which was to be recalled when prompted with the cue word. Critically, each target word was associated with two cue words at the time of initial study (e.g., the word "Horse" could serve as the target word in two different pairs, "Desk – Horse" and "Plane – Horse"). Before participants were tested on their ability to recall each target word, they practiced some of the targets again, either by restudying or by taking a practice cued recall test. However, participants were only given practice with one of the target's two cues (e.g., restudy "Desk – Horse", but not "Plane – Horse"). On the final test, participants were prompted to recall each target word either with the cue word given during practice, or with the unpracticed cue. The results, displayed in Figure 10, showed that memory accuracy and recall latency were only different from a baseline condition of no practice when the same cue was used during both practice and final test.

Why did the benefits of practice fail to transfer between cues? One simple explanation would be that both restudy of paired associates and cued recall practice

simply strengthen the cue to target association, and this specific association is irrelevant on a final test with a different cue. However, reducing the cause of the testing effect to stronger associations from the cue word to the target word is problematic, because the testing effect is strongest following free recall

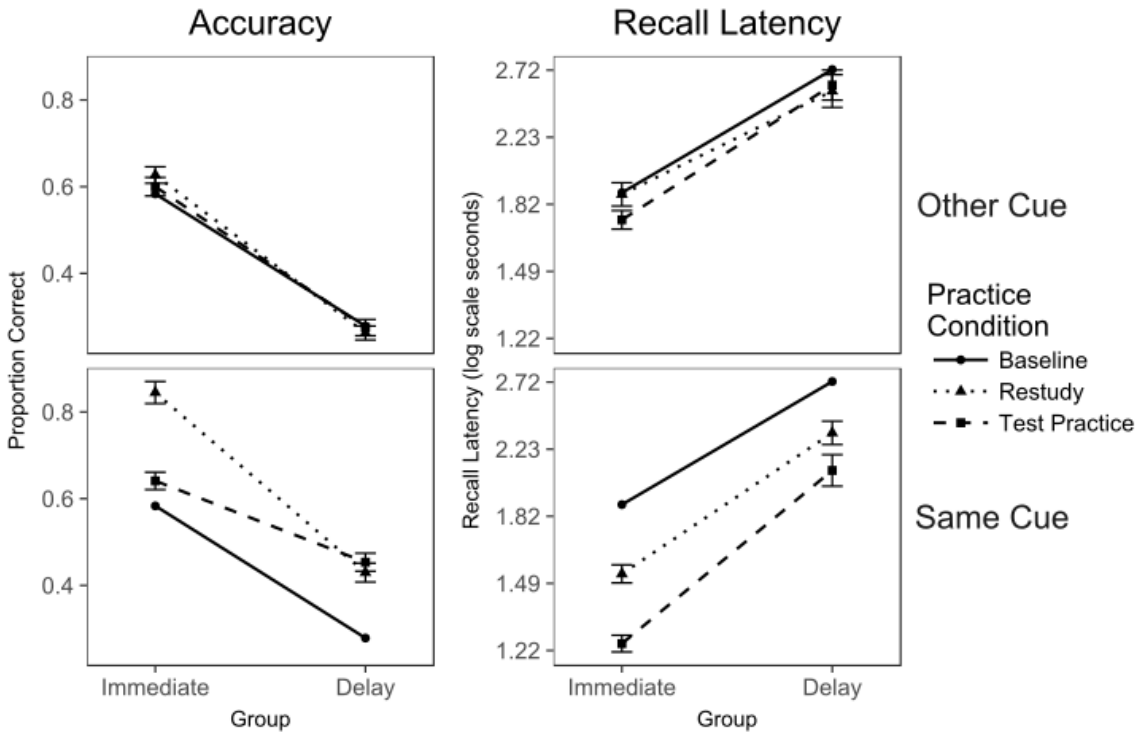

Figure 10: Cued recall memory accuracy and recall latency results from Hopper and Huber (2018), Experiment 2. Error bars represent +/- 1 SEM of the difference from the baseline condition. Note that the baseline condition is duplicated across both rows of the figure in order to enable easier comparison between baseline and each experimental condition.

practice, a test format that specifically lacks explicit cues. The PCR model explains the lack of transfer between retrieval cues as a result of the two different retrieval cues activating two unique sets of item features. As mentioned previously, the PCR model assumes recall strengthens directional associations between the initially retrieved features (in primary retrieval) and subsequently retrieved features (in convergent retrieval). A

consequence of this property is that intra-item associations learned with one cue can fail to transfer to another cue. Because each cue is assumed to activate a set of features unique to the manner in which the target item is considered during encoding, each cue should also utilize a unique convergent retrieval pathway. However, strengthening one particular pathway need not strengthen another. For instance, compare the outcome of the retrieval attempt using Cue X in Figure 2B and 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 produces a different set of initially active features on the target than does Cue X. Thus, strengthened intra-item associations are only beneficial if the set of features activated in primary retrieval is consistent across retrieval attempts.

3.2 Overlap between Semantic Cues

If it were possible for two different retrieval cues to activate the same set of item features, or a highly overlapping set of features, then generalization between retrieval cues would be possible according to the PCR model. Engineering such a situation in practice is challenging, because of the abstract nature of feature representations in the PCR model. Following the assumptions of previous models (e.g., Raaijmakers & Shiffrin, 1981), the features of a to-be-retrieved item are presumed to include perceptual, semantic, phonological, or orthographic attributes of the episode. If these attributes are static (i.e., largely constant across different contexts and pairings), then the features of the memory representation should also be static, making the prospect for being able to engineer a situation where two retrieval cues activate overlapping features good. One possible technique to achieve this goal would be practicing retrieval using a second cue

word semantically related to the target word, in an effort to activate semantic features that would be common across each encounter with a target.

The idea behind this technique is that semantic features may be encoded and in activated any situation. For example, if you needed to associate the unrelated words "Window" and "Dog" in a memory experiment, you might imagine a scenario that links elements of the two words together in some way, like imagining your dog looking at you sadly through your window as you leave for work (see Figure 11). Elements of your semantic knowledge about dogs would be used in generating this scenario (e.g. that they have strong bonds with their owners and do not enjoy long periods of time apart). Thus, one reasonable assumption is that generating this unique association between windows and dogs during study would help the cue word "Window" become associated with the semantic features of the item "Dog" in your memory. Among the set of words that could serve as another retrieval cue for "Dog", a semantic associate should provide the best chance of activating the same semantic information again. For example, if you are prompted to recall a word with the cue "Puppy", that cue will activate a set of features associated with your knowledge about dogs, and this set is likely to overlap with the features considered when encoding "Dog" with the unrelated cue "Window". If recall of "Dog" is successful given the semantic cue "Puppy", and intra-item associations are strengthened, the fact that the features activated in primary retrieval by both "Puppy" and "Window" are likely to overlap means that the intra-item associations just learned are

likely to benefit recall on the final test using the cue “Window”.

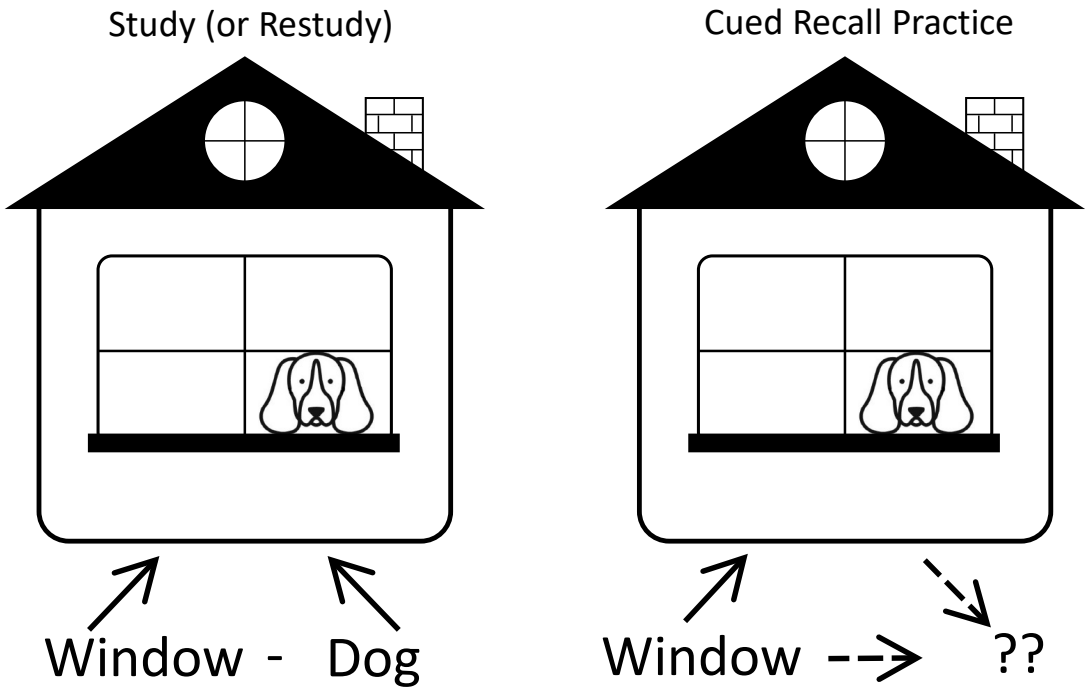

Figure 11: Illustration of different directed associations that are learned when studying (or restudying) the word pair ‘Window – Dog (solid arrows) or when successfully recalling ‘Dog’ in response to the cue word ‘Window’ (dashed arrows). During initial study, a participant might create a mental image of a dog looking out a window, producing directed associations from these words to the mental image. Successful cued recall practice involves activation of the mental image in response to ‘Window and then activation of ‘Dog’ in response to this mental image. According to the PCR memory model learning rule, this strengthens directed associations from this mental image to the word ‘Dog’ as well as directed associations from ‘Window’ to ‘Dog’.

However, it is entirely possible that the set of features used to encode the cue-target pair are *not* largely static. Instead, the features of the encoded memory may include conjunctive attributes corresponding to the content as a pair, and the specific manner in which the pair is interpreted during study (see Criss & Shiffrin, 2004 for a similar proposal). Continuing the “Window” – “Dog” example, the novel conjunction formed (the image of your dog looking at you through a window) may be directly incorporated into the episodic memory trace, along with the features of the individual items.

Importantly, the features representing this conjunction would be independent of the features that might arise to represent a conjunction of, say, “Window” and “Bank”, or “Desk” and “Dog”. In this situation, a reasonable assumption is that although the features of the individual items are well-known prior to the experiment (e.g., both “Dog” and “Window” are well known), the particular conjunction is novel, and the pathways necessary for retrieval of “Dog” from “Window” may not be well-practiced. Thus, you may rely on using the cue word given on the test to activate the features representing the novel conjunction of the words, to in turn activate the features of the correct target word. In other words, the features representing the novel episodic conjunction of the cue and target formed during encoding are mediating recall, not associations between the cue and features of the target item itself.

If features representing the novel episodic conjunction of the word pair form part of the memory trace, and are the features activated by the cue in primary retrieval, then it is likely that intra-item learning stemming from recall with practice another cue (no matter how related to the target) will fail to generalize back to the original cue. On the other hand, if associations between the cue and specific features of the target item itself (e.g., semantic features) are formed during study and are later used during primary retrieval, then it should be possible to observe generalization from recall practice with another cue than can activate the same features. Thus, examining semantic cue generalization in Experiment 2 provides a method of understanding some characteristics of the feature representations stored in memory and used for retrieval: are they static across episodes, representing just individual item attributes along with associations, or

are they dynamic, incorporating features to represent the idiosyncratic conjunctions unique to particular pairings and episodes?

I performed a pilot study to investigate the possibility of transfer of learning from semantically related retrieval cues to unrelated cues. Participants studied lists of unrelated cue-target word pairs, and then immediately received either three rounds of restudy, cued recall practice, or no practice for each target, and took an immediate final cued recall test using the cues learned during initial study. The cue words given during practice were strong forward semantic associates of the target words (e.g., for a target word "Doctor", the cues "Physician" "Nurse", and "Medicine" were used). Practice with these semantically associated cues is analogous to the "Other Cue" condition from Hopper and Huber (2018), only using a semantic cue rather than another independent episodic cue. Multiple rounds of practice with strong semantic associates were used because 1) using multiple strong cues would maximize recall during practice, even without exposure to the cue during the initial study phase, and 2) retrieving the target in response to many different retrieval cues would strengthen intra-item associations from a variety of possible primary retrieval starting points, thus increasing the chance of overlap between the features activated by the unrelated cue word and at least one of the semantically associated retrieval cues.

The results of this pilot experiment showed generalization from cued recall practice with semantically related cues to the unrelated cues provided on the immediate final test (i.e., performance in the cued recall practice condition was above performance in the baseline no-practice condition). However, restudy practice produced an equivalent degree of transfer of learning between cues. A more fine-grained analysis of the practice

test data showed that final test performance was enhanced as long as the target was recalled in response to at least *one* of the three semantically related cues, but that recalling the target in response to two or three of the cues confer no additional performance benefits. Finding that restudy and cued recall practice *both* produced a transfer of learning between cues was a different result than previously obtained by Hopper and Huber; this prior study used only unrelated cues, where *neither* restudy or cued recall practice produced a transfer of learning between retrieval cues. Furthermore, the fact that restudy and test practice produced similar benefits suggests than semantic cue practice presents the opportunity for learning other associations then just intra-item associations. To further understand when practice with one retrieval cue transfers mnemonic benefits to others cues, Experiment 2 extends this pilot study to include several additional conditions. Instead of using a semantic cue for all practiced targets, half of practiced targets will be practiced with their original episodically associated (but semantically unrelated) cue word, and the other half will be practiced using a sematic cue. Furthermore, this experiment will also include a delayed final test condition; if transfer of learning from semantic retrieval cues to the unrelated retrieval cues occurs following test practice *and* is resistant to forgetting after a delay compared to restudy this would suggest that that the transfer of learning between cues stems from strengthened intra-item associations.

3.3 Summary of Experiment 2

Prior research has shown that practice recalling information given one cue does not enhance the ability to recall that same information given another random cue. The PCR model explains the lack of transfer between retrieval cues as a result of these cues

activating two unique sets of item features, reducing the mnemonic contributions of any intra-item associations learned from recall practice. However, if the retrieval cues used during the practice and final tests activate at least a partially overlapping set of target item features, then transfer of learning between cues should be possible. In Experiment 2, this overlap is sought be by practicing retrieval of target words from a cue-target pair learned during initial study using different cue words that are semantically related to the target word. This experiment will develop the PCR model by shedding light on the static or dynamic nature the feature representations stored in memory and used for retrieval.

3.4 Methods

3.4.1 Participants.

94 participants from the University of Massachusetts Amherst were recruited from the undergraduate subject pool. Participants were randomly assigned to either the immediate or delayed final test condition when the experiment commenced. As a result, 48 participants were assigned to the immediate final test condition, and 46 participants were assigned to the delayed final test condition. Participants were actively recruited until at least 45 individuals had participated in each retention interval condition. Once this goal was reached, recruitment ceased, but individuals with scheduled appointments were still allowed to participate. This planned sample size of 45 participants in each retention interval condition was chosen as to be comparable to the previous literature, and specifically Experiment 2 of Hopper and Huber (2018). Participants were compensated with two units of credit they could apply towards extra credit opportunities in undergraduate psychology classes.

3.4.2 Materials.

120 English word triplets (consisting of a target word, a semantically related cue word and an unrelated cue word) were generated using a two-step process. In the first step, pairs of target words and semantically related cue words meeting specific criteria were selected from the USF word association norms database (D. L. Nelson, McEvoy, & Schreiber, 2004). The specific criteria were:

1. All target words must have between four and 10 letters, and a word frequency between five and 200 uses per million words according to the $SUBTLX_{US}$ corpus (Brysbaert & New, 2009).
2. Each cue has a minimum forward associative strength of 0.2 to its target word[9].
3. Each word in the combined set of cue words and target words is unique.
4. Words use either their plural or singular form, but never both (e.g., "horse" and "horses" cannot both appear in the combined set of cues and targets).

The top 120 word pairs with the highest forward associative strength between the cue word and the target word were chosen from this set. In the second step, 120 new words with a forward associative strength of zero to all the chosen targets were selected. Each of these 120 unrelated words will serve as the episodically associated cue for a randomly chosen target word, thus completing each triplet. As an example, a possible triplet could be "Toad" (the target), "Frog" (the semantic cue), and "Wall" (the unrelated cue). The same 120 word triplets were used as the stimulus materials for each subject. The word triplets were randomly divided into six lists of 20 for each participant. The experiment

[9] The forward associative strength was determined using a free association task where participants were provided with cue words, and asked them to name all the words that come to mind in response to each cue. The value of the forward associative strength between a cue word and a response word was calculated the proportion of times that response was given to that specific cue word. For example, the cue word "bride" elicited the response "groom" 86.5% of the time, giving bride a forward associative strength to "groom" of .865. For more information, visit w3.usf.edu/FreeAssociation/Intro.html

was performed entirely on a desktop computer using MATLAB and the Psychophysics Toolbox.

3.4.3 Procedure.

A diagram outlining the entire experimental procedure for both groups of participants is shown below in Figure 12. Participants completed an initial study phase followed by a practice phase and a final test phase for each of the six lists. In the initial study phase, participants studied each individual target word together with its unrelated cue word for five seconds. The unrelated cue word and target word from each triplet were presented on the left and right sides of the screen, respectively (e.g., Wall – Toad). Participants were instructed to learn these pairings, and told they would need to recall the word on the right (the target word) when shown only the word on the left (the cue word) on a future test.

The practice phase immediately followed the initial study phase for each list. Target words were randomly assigned to one of five practice conditions: unrelated cue test practice (UC-TP), unrelated cue restudy (UC-RS), semantic cue test practice (SC-TP), semantic cue restudy (SC-RS), or no practice (control). The unrelated cue practice conditions may also be referred to as episodic cue practice conditions (the label "Unrelated Cue" describes how the cue word was chosen in the stimulus pool, while the label "Episodic Cue" describes how it was associated with the target word during the experiment). For target items assigned to receive test practice, participants were shown a cue word, and given 10 seconds to recall the correct target word from the initial study phase (i.e., cued recall practice). For target items in the unrelated cue test practice condition, the cue word shown was the same word seen with the target immediately prior

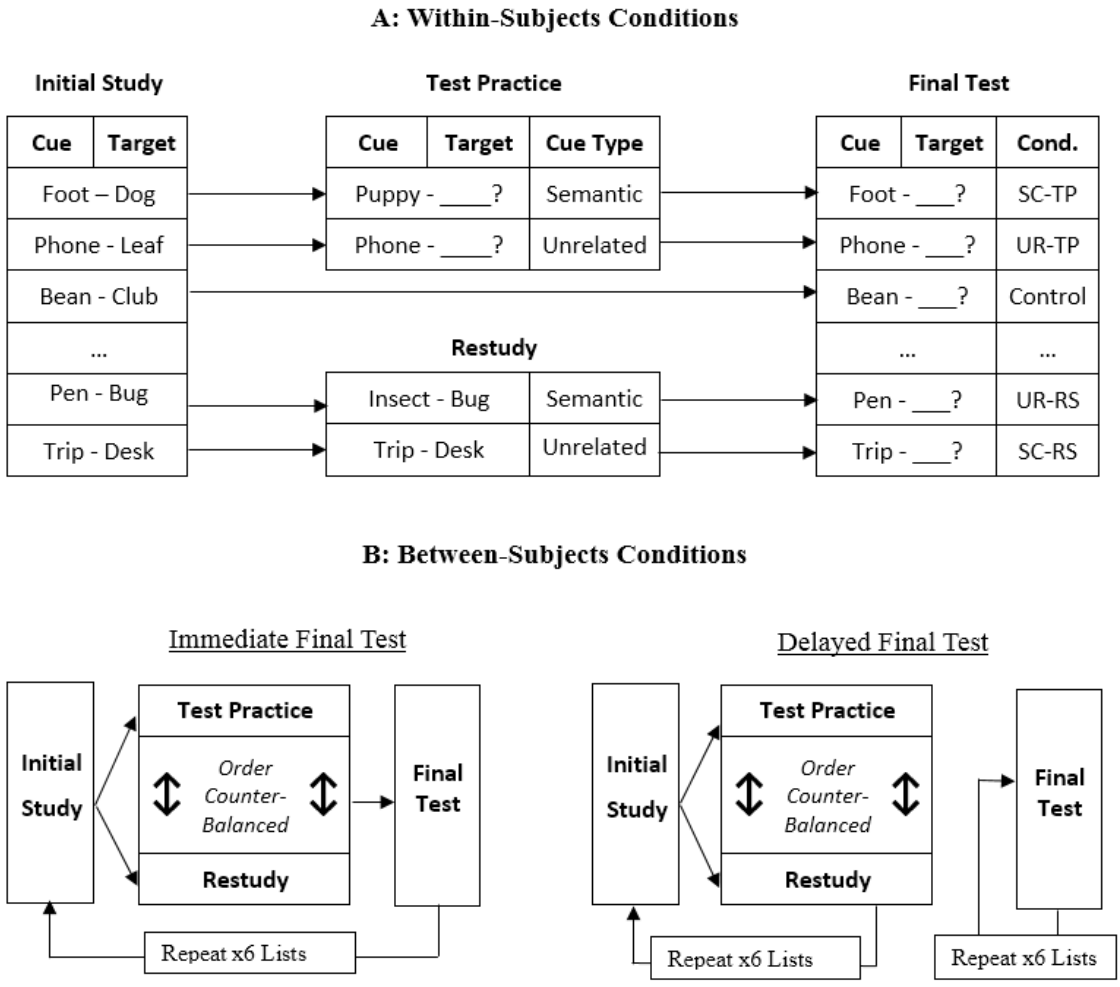

Figure 12: Schematic diagram of the procedure in Experiment 2. The upper portion of the figure depicts the methods of initial study, practice, and final testing for target items in each of the five practice conditions. The lower portion of the figure depicts the difference in final test timing between the immediate and delayed final test trials.

in the initial study phase (e.g., Wall – ____?). For target items in the semantic cue test practice condition, the cue word shown was the semantically related word selected from the USF word association norms (e.g., Frog – ____?). For target items assigned to be restudied, participants were shown a cue word on the left and the target word on the right for five seconds, just as in the initial study phase. For target items in the unrelated cue restudy condition, the cue word was the same word seen with the target immediately prior in the initial study phase (e.g., Wall – Toad). For target items in the semantic cue

restudy condition, the cue word was the semantically related word selected from the USF word association norms (e.g., Frog – Toad).

To help participants perform the correct retrieval task on each test trial (i.e., to recall the target word seen with the cue during the initial study phase, or recall a target word from the study list that is semantically related to the cue), the cue words were color-coded during the practice phase. Cue words semantically related to the target word were colored green, and cue words unrelated to the target word (but paired with the target during initial study to form an episodic association) were colored red. These color codes were also used on restudy trials for consistency. Participants gave their responses on test trials with the computer keyboard, and were permitted to edit their responses on test trials by using the Backspace key before confirming them with the Enter key. No feedback was given about performance on practice test trials. Restudy trials and cued recall test practice trials were grouped into separate blocks of trials during the practice phase, and the order of these blocks was counterbalanced across subjects. The presentation order of the word pairs was randomized within each list, so that words were not practiced in the same order as during the initial study phase.

After the practice phase, each target was given a final cued recall test using the unrelated cue word from the initial study phase as a probe. The timing of the final test varied between the immediate and delayed final test conditions. Participants in the immediate final test condition completed the final test for a given list immediately after practice for that list. Thus, they had six rounds of initial study, practice, and final test, in that order (e.g., study-practice-test for list one, followed by study-practice-test for list two, etc.). In the delayed final test condition, the final test will not occur until after all six

lists receive initial study and practice. Thus, participants had six rounds of initial study and practice, and then one long final test covering all six 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 final test for each list in the delay condition was given in the same order the lists were studied in (i.e., the targets from list one were tested first, then the targets from list two, etc.).

As during the practice tests, participants were given 10 seconds to type in the missing target word using the computer's keyboard, and were permitted to edit their responses by using the Backspace key before confirming them with the Enter key. The order in which the word pairs were tested was randomized within each list, so that the order of pairs on the final test will not be the same as in the initial study or practice phases. A 30 second break was given in between the final test for each list (i.e., after every 20 trials). The entire experiment lasted approximately 45 minutes.

3.5 Results

3.5.1 Scoring.

Recall latencies were measured as the duration between the onset of the cue word, and the first key press of the word entered as the participant's final response. The accuracy of each response was scored by a software routine that allowed for small misspellings (e.g., letter transposition, pluralization) to be labeled as correct.

3.5.2 Statistical Analysis.

Both the recall accuracy and latency data were analyzed using hierarchical Bayesian regression models. A logistic mixed-effects model was fit to the recall accuracy

data, and a linear mixed-effects model was fit to the recall latency data after log transforming the latencies. In both cases, practice type (episodic cue restudy, episodic cue test practice, semantic cue restudy, sematic cue test practice cued recall, and no practice) and retention interval (immediate vs. delayed final test), as well as their interaction, were included as fixed effects. The models also included a random intercept term for each participant. The *rstanarm* package for the R computing language (R Core Team, 2017; Stan Development Team, 2016) was used to estimate the posterior distributions of both models.

A Gaussian probability distribution was used as the prior distribution for all regression coefficients in both models. The **μ** and **σ** parameters for these prior distribution were informed by estimates from two previous studies. The parameters for the prior distributions on the intercept, retention interval effect, episodic cue practice effects, and control condition were based on a mixed-effects model analysis of the data from Hopper and Huber (2018). The parameters for the semantic cue practice main effects were based on a mixed-effects model analysis of data from the pilot study described in the introductory section of this experiment. The estimated fixed effect coefficient values from these models were then used as the values for the **μ** parameters of the prior distributions in the current analysis. The standard errors of the fixed effect coefficient were used as the **σ** parameters, after multiplying by three to instantiate the assumption of additional uncertainty owing to a slightly different experimental design, stimuli set, and new sample of participants. Because neither of the previous experiments included a retention interval manipulation along with semantic cue practice (and thus, could not measure interactions between them), the default weakly informative prior distribution

was used for the semantic cue by retention interval interaction coefficients. The specific parameter values used for each coefficient's prior distribution are reported in Table 3 and Table 7, for the accuracy and latency models respectively.[10]

Contrasts of the estimated recall accuracy and latency between conditions were performed using the *emmeans* R package (Lenth, 2018). Contrasts were computed by subtracting the distributions of posterior predicted values between conditions (on the log-odds and log scale respectively), and examining whether the 95% Highest Density Interval (HDI) of this distribution contained zero. If zero was outside the HDI, the contrast was interpreted as evidence for a reliable difference.

The observed proportion of correctly recalled words in each condition is shown in the top panel of Figure 13, and the proportion of correctly recalled words in each condition estimated by the model is reported in Table 4. First, accuracy differences between practice conditions were assessed at each retention interval (see Table 5 for a numerical summary of all practice condition contrasts). All practice conditions produced higher recall accuracy than the baseline no practice condition, at both retention intervals. On the immediate final test, episodic cue restudy produced the highest recall accuracy (Mdn = 79.6%), reliably greater than restudy with a semantically related word (Mdn = 63.1%) and test practice with the same episodically related cue word provided on the final test (Mdn = 55.8%). Semantic cue restudy, which produced the second highest accuracy on the immediate final test, was reliably greater than accuracy in both the

[10] While using informed priors is always advisable, it made little difference in this case. The models described in this section led to the same qualitative conclusions as models with the same structure but using the default weakly informative prior distribution $N(0, 2.5)$ for all parameters.

episodic and semantic cue test practice condition (Mdn = 57.0%). The two forms of test practice (semantic cue and episodic cue) produced nearly identical performance on the final test.

Overall recall accuracy dropped sharply over the retention interval, but the rank order of the practice conditions on the delayed final test was mostly similar to that of the immediate final test. The episodic cue test practice condition was the single exception, which now showed the highest recall accuracy of all practice conditions. Importantly, the relationship between the episodic cue restudy and the episodic cue test practice conditions reversed, with episodic cue test practice now producing higher accuracy (Mdn = 39.2%) than episodic cue restudy (Mdn =37.9%). This reliable cross-over interaction reflects the classic "testing effect" (see Table 6 for a numerical summary of the interaction contrasts). Thus, testing produced superior retention relative to restudy, at least when using the same retrieval cues during both practice and final testing.

Further interaction contrasts demonstrate that the retention benefits from test practice were highly specific. Unlike episodic cue restudy and test practice, no interaction was found involving the semantic cue restudy and semantic cue test practice conditions. Instead, accuracy in the semantic cue restudy condition (Mdn = 63.1% for the immediate final test, Mdn = 29.5% for the delayed final test) was consistently greater than in the semantic cue test practice condition (Mdn = 57.0% on the immediate final test, Mdn = 26.1% on the delayed final test). A direct comparison between the two types of test practice (episodic cue and semantic cue) further demonstrate their differential effects. Semantic cue and episodic cue test practice produced nearly identical recall accuracy on

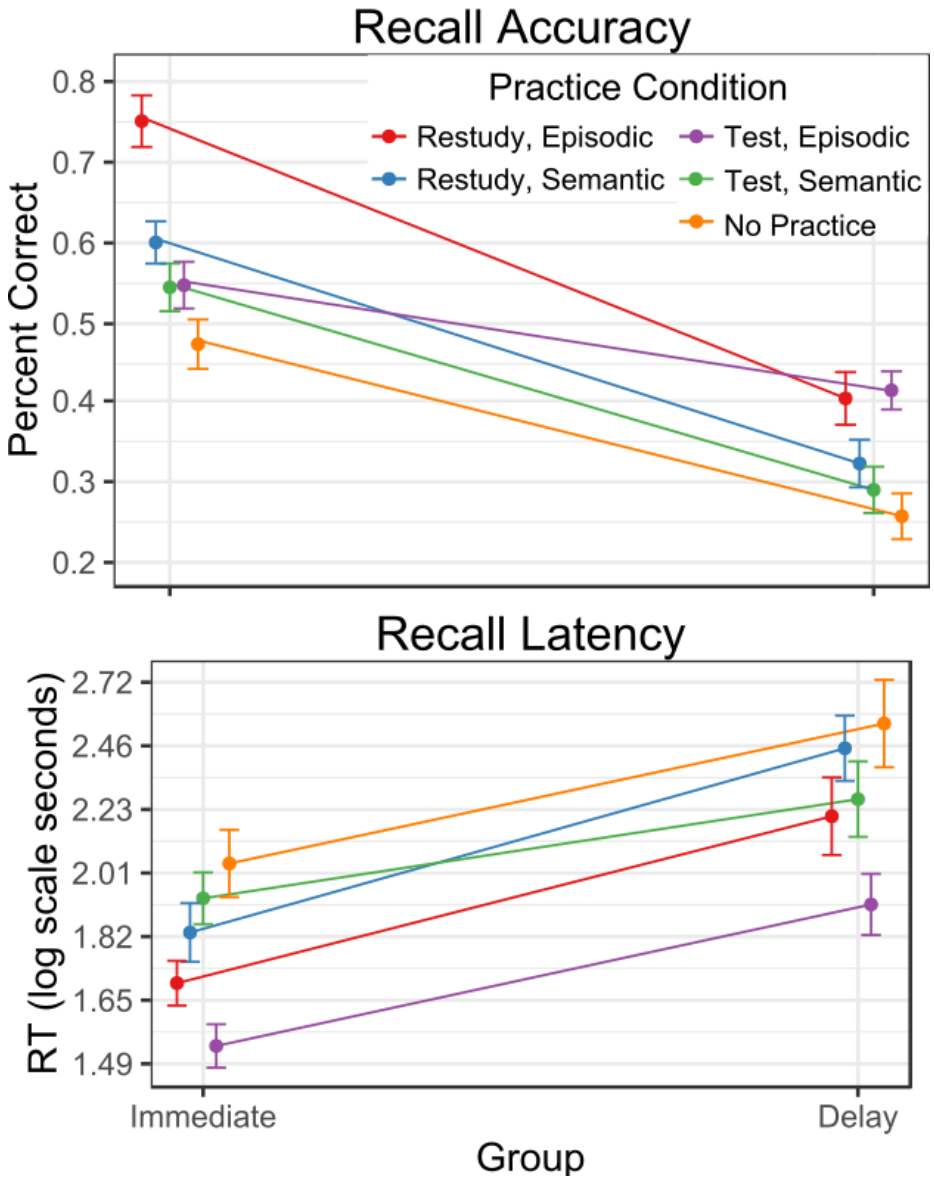

Figure 13: Recall accuracy and latency from Experiment 2. Error bars represent 95% confidence intervals calculated using the subject-normalized method of Morrey (2008).

the immediate final test (Mdn = 57.0% and Mdn = 55.8%, respectively), but they greatly diverged on the delayed final test, with accuracy in the episodic cue test practice condition remaining relatively high (Mdn = 39.2%) and accuracy in the semantic cue test practice condition dropping down more sharply (Mdn = 26.1%). The divergence of the semantic and episodic cue test conditions, and the lack of cross-over between the two

semantic cue conditions, show that realizing the retention benefits of test practice depends on using the same retrieval cues on the final test as during practice.

The average recall latency for correctly recalled words in each condition is shown in the bottom panel of Figure 13, and the median recall latency in each condition estimated by the model is reported in Table 8. First, recall latencies were compared across practice conditions at each retention interval. Mirroring the recall accuracy results, active practice produced faster recall than the baseline no practice condition, at both retention intervals. However, practice conditions that produced more accurate recall did not always produce faster recall. Episodic cue test practice produced the fastest recall (Mdn = 1.51s), followed by episodic cue restudy (Mdn = 1.66s), semantic cue restudy (Mdn = 1.78s) and semantic cue test practice (Mdn = 1.85) in turn. All pairs of these conditions were reliably different, with the exception of the semantic cue test practice and semantic cue restudy conditions.

Correct recall latency slowed over the retention interval, but the pattern of latencies across conditions on the delayed final test was similar to the immediate final test. Again, the episodic cue test practice produced the fastest recall (Mdn = 1.96s), followed by episodic cue restudy (Mdn = 2.17). The semantic cue restudy and semantic cue test practice conditions reversed rank order at the delayed final test, with a median recall latency of 2.26s for semantic cue test practice and a median recall latency of 2.36s for semantic cue restudy. The interaction between the two semantic cue practice conditions and the retention interval narrowly reached the established criterion for

reliability, with the HDI of the interaction contrast excluding zero by a margin of .001. No other interactions between practice types and retention interval occurred[11].

3.6 Discussion.

All tests do not produce the same learning – this is the certain conclusion from Experiment 2. In terms of final test recall accuracy, retention was only enhanced when the same retrieval cue was provided for both the practice test and final test. This can be clearly seen in Figure 13, where only the episodic cue test practice condition shows a different slope across the immediate and delayed final tests. When a word semantically related to the target word was given as a retrieval cue on the practice test, forgetting over the course of the retention interval followed a similar trend as both restudy conditions. The consistency of the advantage for restudy over testing when practicing with the semantically related cue (rather than given the same cue provided on the final test) indicates that consistent retrieval cues are necessary to realize the retention benefits of test practice does. In other words, it appears the retention benefits of retrieval do not generalize across item-level retrieval cues.

The conclusions from the analysis of the recall latency data were similar. Not surprisingly, recall was faster whenever the same cue word was provided during the practice phase and final test phases. Among these episodic cue conditions, test practice produced faster recall than restudy at both retention intervals. No such consistency was seen among the semantic cue practice conditions, and in fact, the latency relationship between the semantic cue restudy and semantic cue test practice conditions reversed over

[11] For completeness, Table 10 summarizes the recall latency interaction contrasts examined.

the retention interval. However, there is no clear theoretical explanation reason for such an interaction, and an interaction involving recall latency changes over time has not been seen any previous published studied (Hopper & Huber, 2018; van den Broek et al., 2014). Because of this, and because of this interaction's extremely narrow margin of significance, I believe it is prudent to regard it as spurious. Thus, semantic cue practice (restudy or testing) do not appear to have to produce different impacts on recall latency when the final test uses another item-level cue.

These results replicate the findings of Hopper and Huber (2018), further bolstering the conclusion that learning from tests is cue-specific in practice. The mechanisms of the PCR model allow for generalization across item-level retrieval cues to occur in theory, but achieving this generalization requires that both retrieval cues are associated with overlapping sets of item features, and this overlap appears difficult to achieve. These results, along with the earlier results of Hopper and Huber (2018), are consistent with the idea that memory representations for episodes with multiple items are more than the sum of their individual features. Unique or "emergent" features that are specific the conjunction of the items, and independent of the features defining their individual constituents, have been used to explain differences between single item and associative recognition in computational models of memory (Eich, 1985; Murdock, 1982, 1993). More recently, this assumption of emergent associative information has been included in an extension of the REM model (Criss & Shiffrin, 2004, 2005) to account for a lack of interference in associative recognition between pairs of items that shared one member in common, but whose second member came from a different class of material (e.g., studying the word "Traffic" paired with the image of a face did not impair

recognition of a word-word pair like “Traffic” – “Oxen”). Criss and Shiffrin (2005) also demonstrated that single item recognition performance for an item that was repeated across pairs did *not* depend on the type of pair the item belonged to (i.e., repeating an item across two word-word pairs did not differ from when the item was repeated across a word-word pair and a word-face pair), indicating the some stable, item-specific information is also encoded along with the emergent pair representation.

The use of either emergent pair-specific features, or stable item-specific features, to drive retrieval can explain the lack of generalization observed across semantic and episodic cues in Experiment 2. While the episodically associated cues relied on activating the emergent associative features as the starting point in primary retrieval, the semantically related cues may have relied on activating the stable, item-specific features, to enable retrieval of the target, because these are the semantic features they share in common with the target word. Because the pair-specific features and the item-specific features represent distinct groups, convergent retrieval learning beginning from one set of features will not benefit the convergent retrieval process when it begins from another set. Thus, while the results of Experiment 2 represents “null” results in terms of cue generalization, they provide guidance for theory and model development. Future computational implementations of the PCR model should incorporate the assumption of both emergent pair-specific features and stable item-specific features when encoding multiple items into a memory representation.

Presumably, the semantic cues also activated the features of other items in memory that were not among the words initially studied (e.g., “Puppy” will also activate semantically related competitors to “Dog”, like “Kitten”). Paring down any activated

semantic competitors to enable retrieval of the correct item from the studied list would require the joint use of a temporal context cue. Using a temporal context cue to filter the list of semantic associates activated on the practice test may also explain why recall was slightly above baseline in both the semantic cue test practice conditions of this experiment, while recall was equal to the baseline condition in both the analogous "Other Cue" conditions of Hopper and Huber (2018). In this prior study, using a temporal context cue would not have been beneficial to performance, as the temporal context cue would be equally associated with the correct target word as with all the other competing words from the list. In that situation, using the item-specific cue would provide the most discriminative information. While it is true that the temporal context cue would be equally associated with the target word (i.e., "Dog") as with all the other competing words from the studied list in the current experiment as well, in this situation it proves extremely useful in *conjunction* with the semantic cue; the target with an association to both the semantic cue *and* the temporal context cue will always be the correct one to recall. If participants did use information from the temporal context jointly with the semantic cue in the practice phase of Experiment 2, this would produce a strengthened association with the temporal context cue information and any retrieved item. Because of the (at least partial) consistency between temporal context at both the immediate and final tests, having a strengthened association between the temporal context cue and the target item would improve retrieval on the final test relative to baseline (because baseline items have no opportunity for such additional learning during the practice phase).

A competing theoretical account of retrieval-based learning proposed by Rickard and Pan (2017) provides an alternative explanation for the cue-specific benefits of test

practice observed here and elsewhere (Hopper & Huber, 2018; Pan et al., 2016; Pan, Wong, Potter, Mejia, & Rickard, 2015). The claim of the "Dual Memory" theory is that restudy and test strengthen different memory traces; restudy is assumed to strengthen the memory trace formed during initial study, while retrieval both strengthens the initial trace (on which practice test retrieval is based) *and* encodes a new, separate test memory trace. This account can be seen a more specific version of the Bifurcation model (i.e., the reason the total memory strength increases more for retrieved items than studied items is because retrieved items can draw on the strength of two memories instead of one). The new test trial memory has two components: cue memory, an episodic encoding of the retrieval cue in a retrieval context, and an association between this cue memory and the correct response. Thus, any learning about the correct response is cue-specific. The Dual Memory account, which assumes retrieval encodes a new memory trace, directly conflicts with the PCR model, which assumes retrieval strengthens an existing memory trace. It is difficult to discriminate between multiple and single trace theories (as they often make similar predictions) but as developed further in the general discussion, the Dual Memory and PCR accounts of retrieval based learning may be able to be contrasted by examining the effects of retrieval on other memory effects, such as the list-strength effect in recognition.

CHAPTER 4

EXPERIMENT 3

4.1 Visually Guided Convergent Retreival

Under the PCR model, recall enhances retention and subsequent recall latency relative to restudy opportunities because an item's features gradually become active in a staged manner during retrieval, resulting in stronger associations between these features that can be leveraged during future retrieval attempts. The intra-item learning that putatively takes place following recall is an emergent property of the model's assumption that retrieval cues partially activate items in memory during recall attempts, and that directional associations are strengthened between features when they become active in sequence. However, learning associations between the features of an item is not necessarily limited to taking place during recall; the circumstances of a recall tests merely present an opportunity to do so. A unique prediction of the PCR model is that if strong associations between the item features necessary for recall can be formed during encoding, retention and recall latency should be enhanced relative to traditional study, even without engaging in an explicit recall task.

Experiment 3 tests this prediction using visual stimuli, which have more concretely definable features than purely verbal stimuli. In Experiment 3, participants studied images of common objects by viewing the whole object, or being shown an accumulation of individual features from the object in rapid sequence. After the learning phase, participants were be tested on their ability to recall the name of each object when prompted with one of the object's features. The PCR model predicts that experiencing the individual features from the object in rapid sequence will lead to stronger associations

between the features than restudying the whole image, promoting better long term retention and faster recall of the details necessary to recall the object.

4.2 Methods

4.2.1 Participants

The participants were 72 individuals from the University of Massachusetts Amherst, all recruited from the undergraduate subject pool. Of the 72 recruited participants, 65 returned for the second session of the experiment. Participants who completed both sessions received two credits that could be applied either toward extra credit opportunities in undergraduate psychology classes, while those who completed just the first session received only one credit

4.2.2 Materials

A total of 120 photographic images of common objects (e.g., a tea kettle, a cardboard box, a bike, a chair) were converted to a greyscale color pallet. All objects were displayed alone in the image against a white background. During the experiment, all images were displayed foveally at size of 512 by 512 pixels. Some images were resized downward to the desired resolution, but no images were enlarged from lower resolutions. The stimuli used in the experiment can be viewed and downloaded from https://github.com/cMAP-CEMNL/VGCR_experiment/tree/master/stimuli.Images were divided up into the cells of a three by three grid, thus defining nine visual features for each object. An important criteria for selecting the set of images used in this experiment was how much of the 512 by 512 image was taken up by the object, and how evenly distributed the object was throughout the 512 by 512 image. Specifically, it was required

that the object occupy at least 25% of the pixels in each cell. For example, an object such as a hockey stick or a sword would be eliminated using this criteria. This requirement serves to minimize heterogeneity between the visual information provided by the content of each cell of a given object (e.g., preventing the object from being entirely revealed by three of the cells, while the remaining six are empty).

4.2.3 Procedure

Each of the 120 objects was randomly assigned to one of three viewing conditions: whole object study, "guided convergent retrieval" (GCR) practice, or no practice (to serve as a baseline control condition). In the whole object study condition, the objects were presented on the screen for 2.25 seconds. In the GCR condition, objects appeared on the screen gradually over time in a staged fashion, with a total presentation time of 2.25 seconds also. After presenting the object, the name of the object was presented alone on screen for two seconds in both the whole object and GCR conditions.

In the GCR condition, the object was revealed gradually by dividing it into nine equally sized, non-overlapping portions (i.e., a three by three grid) that together covered the entirety of the image (these portions will hereafter be referred to as "features"). These object features were shown one at a time, in the same position on the screen it would occupy if the complete object were being displayed. The order in which each object feature appeared in the GCR condition was intentionally specific. The first feature to appear was selected based on data from a stimuli naming pilot study, where participants viewed three features from each of the 120 objects, and attempted to name the object the feature came from each time (having never viewed the entire object itself). For each object, the feature with the lowest naming accuracy in the pilot study was chosen to be

the first feature displayed in the GCR condition. The selection method was important because this feature also served as the retrieval cue during the test phase, and using features with inherently high naming accuracy would potentially introduce ceiling effects that obscure differences between the practice methods. After the first object feature appeared, the next feature to appear was always adjacent. For example, if the top-left feature in the three by three grid was the first feature to appear, the second to appear would either be the top-middle or middle-left feature. This "adjacent feature" rule applied to all subsequent features, and each feature was displayed for 250 ms. before the next feature appeared, for a total trial length of 2.25 seconds. A diagram demonstrating format of the GCR condition is shown Figure 14.

Participants were exposed to the 40 GCR practice items and the 40 whole object study items in random order. After all participants were exposed to all 80 items, they received a second round of exposure to all of the items, in another random order. After the exposure phase, participants were given a cued recall naming test. For half the items in each practice condition, the test phase occurred after a one minute math distractor task (calculating a cumulative sum of 6 numbers between one and nine) following the exposure phase. This condition is hereafter referred to as the immediate test condition. For the remaining items, the test phase occurred in a second session (hereafter referred to as the delayed test condition). The delayed test occurred the day following the exposure phase (approximately 24 hours after the first session). Objects were randomly assigned to the immediate and delayed test conditions.

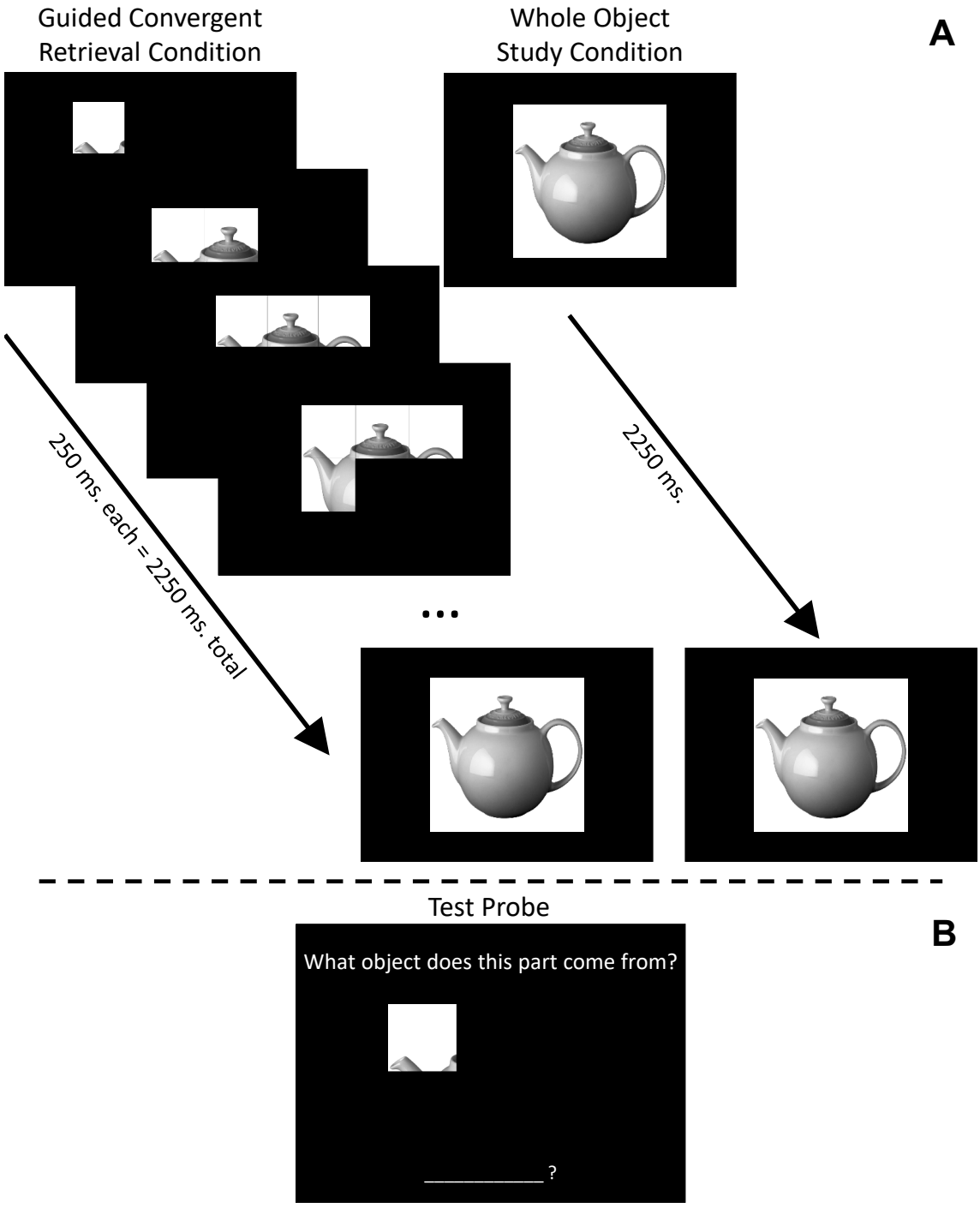

Figure 14: A schematic diagram of the procedure in Experiment 3. Panel A demonstrates the two types of encoding conditions used during the learning phase, and Panel B demonstrates a sample trial from the memory test given following the learning phase.

On the cued recall test (as shown in Panel B of Figure 13), participants were shown a single feature from each object as a cue (i.e., one of the object's nine equally

sized non-overlapping portions). The feature's spatial location was preserved during testing (i.e., it was positioned on the screen in the same location as it would normally occupy if the entire object was presented). Participants responded to the feature cue by typing in the name of the object the feature came from (e.g., typing in "bike" if they remember that the feature shown was part of a bicycle). Participants were permitted to edit their responses by using the Backspace key before confirming them with the Enter key.

4.3 Results.

4.3.1 Scoring

Recall latencies were measured as the duration between the onset of the cue feature, and the first key press of the word entered as the participant's final response. For many of the objects, several synonyms were accepted as correct identifications (e.g., "speakers", "bass speaker" and "amp" were all accepted as possible names for the image showing a large audio speaker). The full list of accepted names can be viewed at github.com/cMAP-CEMNL/VGCR_analysis/blob/master/Stimuli/object_names.csv. The accuracy of each response was scored by a software routine that allowed for small misspellings (e.g., letter transposition, pluralization) to be labeled as correct.

4.3.2 Statistical Analysis

Differences in accuracy and recall latency between conditions were assessed with a mixed-effects regression models, using the *lme4*, *afex*, and *emmeans* packages for the R statistical computing environment. The model of naming accuracy utilized a logistic link

function, while the model of naming latency utilized a logarithmic transformation of the observed response times in to meet the assumption of homoscedastic Gaussian residual variance in the regression model. Both models were fit by minimizing the residual maximum likelihood (REML). Practice type (whole object study, GCR study, and no practice) and retention interval (immediate vs. delayed final test) factors, as well as their interaction, were included as fixed effects in both the accuracy and latency models.

Both models also included random intercepts for each participants and object in the experiment. This random effects structure was reached by starting with the maximal random effects structure (i.e., random intercepts and slopes for both participants and items in each condition), and removing terms from the random effects structure (beginning with the item component) until the model fitting routine was able to converge on stable parameter estimates[12]. For the regression model of accuracy, main effects were assessed using likelihood ratio tests, and contrasts were performed using Holm-Bonferroni corrected Wald tests. For the regression model of latency, main effects were assessed with ANOVA using the Kenward-Roger approximation of the error degrees of freedom, and contrasts were performed using Holm-Bonferroni corrected *t*-tests (also using the Kenward-Roger degrees of freedom approximation).

Figure 15 shows the average naming accuracy observed in each experimental condition. As expected, baseline naming accuracy for unpracticed objects was low,

[12] The naming latency model was able to estimate random slopes for each participant, but not while also include random item effects in the model. On the other hand, the accuracy model was never able to estimate per-participans slopes, regardless of whether item-effects were included in the model. I decided use the naming latency responses model with per-object intercepts over the model with per-participant slopes in order to have comparable structures between latency and accuracy models, and because of the high variability in naming accuracy across different objects.

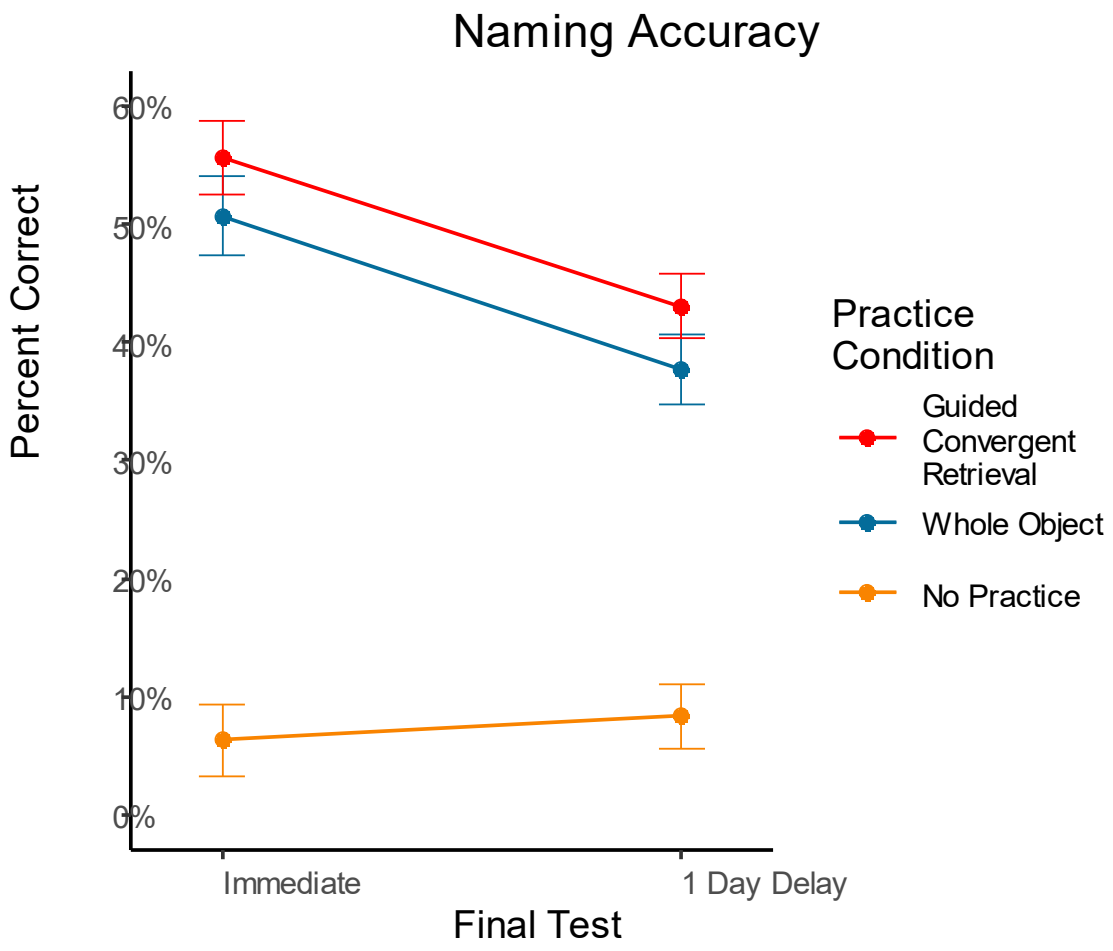

Figure 15: Object naming accuracy in Experiment 3. Error bars represent 95% confidence intervals calculated using the subject-normalized method of Morrey (2008).

averaging approximately 8% across both the immediate and delayed tests. Exposure to the object during the learning phase clearly had a large effect on naming accuracy, with average naming accuracy for practiced objects jumping up to near 50% across both the immediate and delayed tests. Likelihood ratio tests indicated significant main effects of retention interval ($\chi^2(1) = 36.99$, $p < .0001$) and practice method ($\chi^2(2) = 1928.57$, $p < .0001$), as well as an interaction between retention interval and practice method ($\chi^2(2) = 34.57$, $p < .0001$).

The nature of this interaction was investigated using interaction contrasts (i.e., differences of differences between experimental conditions). Reported in Table 11, these contrasts revealed that the interaction was driven by a decrease in naming accuracy for both practice conditions over the retention interval, combined with a minimal increase in naming accuracy for the baseline no practice condition across the retention interval. In other words, the difference between the whole object study condition and the baseline condition decreased over the retention interval, as did the difference between the guided convergent retrieval condition and the baseline condition. This interaction simply demonstrates that forgetting took place for practiced objects, while no forgetting took place for unpracticed objects, which is to be expected: forgetting should only take place when there is initial practice and learning that can be forgotten.

The interaction contrasts also indicated that difference between the whole object study condition and the guided convergent retrieval condition did not change over the course of the retention interval; in other words, there was no interaction between the two methods of practice. Differences between the two conditions were further examined using contrasts collapsing over the effect of retention interval. This contrast showed that guided convergent retrieval practice produced higher accuracy than whole object study ($z = 4.739$, $p < .0001$).

Figure 16 shows the average naming latency observed for trials in each experimental condition where the object's name was successfully recalled. The low naming accuracy for baseline unpracticed items was complemented by extremely long naming latencies. But whereas naming accuracy for baseline unpracticed items was stable across the immediate and delayed final test conditions, naming latency for these items

sped up greatly between the immediate and delayed final tests. Naming latency for practiced items were much faster overall, but showed the opposite pattern over the retention interval, slowing down as time passed between study and test.

The ANOVA indicated a significant main effect of practice condition ($F(2, 2659.45) = 121.43$, $p < .0001$), no main effect of retention interval ($F(1, 2657.7) = 0.01$, $p = .93$), but a significant interaction between practice condition and retention interval ($F(2, 2635.57) = 10.42$, $p < .0001$). Interaction contrasts were performed to investigate the nature of this interaction. As shown in Table 12, this interaction was driven by the convergence of the baseline condition with the two practice conditions: the difference in naming latency between the whole object study condition and the baseline condition decreased over the retention interval, as did the difference between the guided convergent retrieval condition and the baseline condition. On the other hand, the two practice conditions showed the *opposite* qualitative pattern. The whole object study condition and the guided convergent retrieval condition produced nearly identical naming lantencies for objects tested on the immediate test, but naming latency slowed down more over the retention interval for objects in the whole object study condition than the guided convergent retrieval condition. While this diference of differences did not reach the threshold of statistical significance, the data show a qualitatively different relationship between the latencies in each practice condition on each test. Thus, futher comparisons contrasts between the whole object study condition and the guided convergent retrieval condition were performed both using contrasts where estimates were collapsed over the retention interval, as well as with contrasts of the two conditions separately at each retention interval.

When comparing the whole object study condition and the guided convergent retrieval condition at the immediate final test, there is clearly no difference in mean naming latency, a conclusion corroborated by a null contrast ($t(2650.1) = 0.405$, $p = 0.68$). At the delayed final test responses to objects in the whole object study condition were about 8% slower than to objects in the guided convergent retrieval condition, and

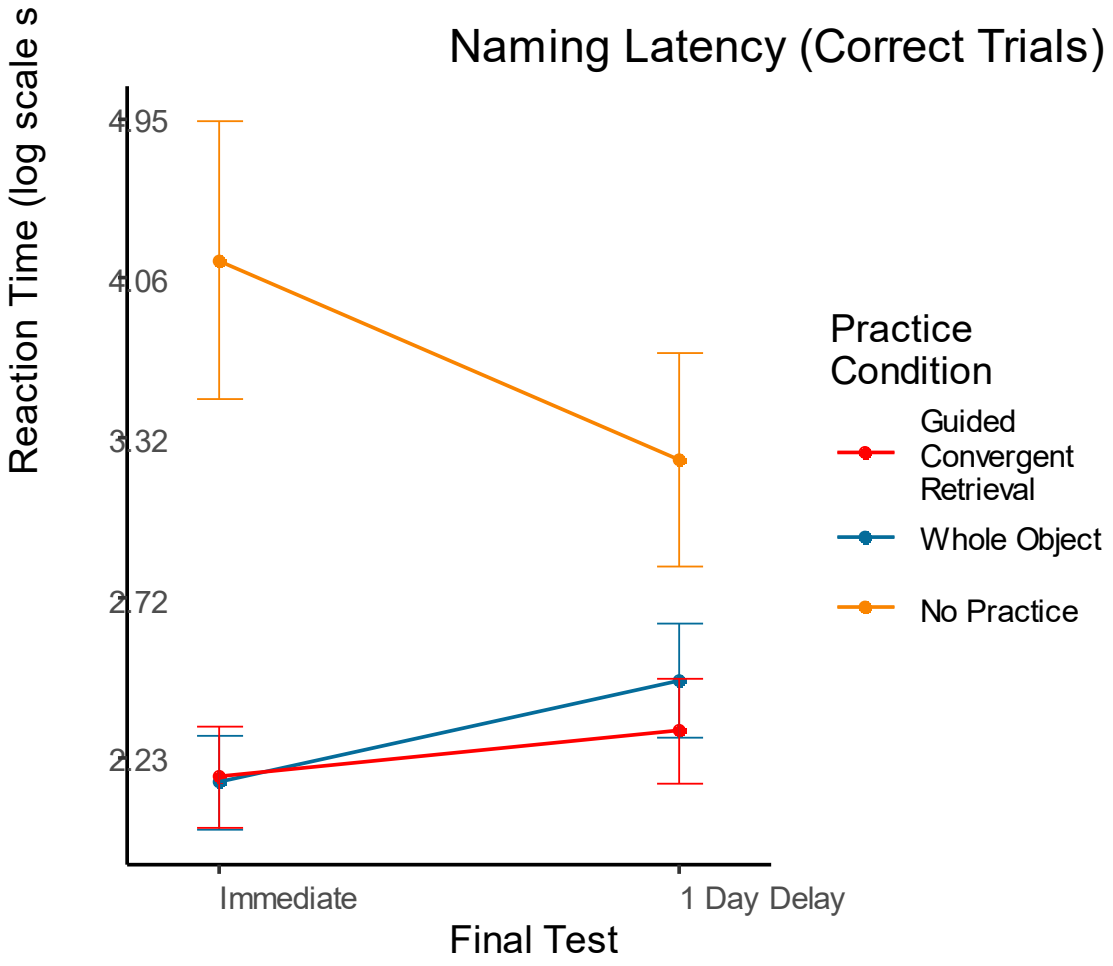

Figure 16: Object naming latency for trials in Experiment 3 where the object was correctly identified. Error bars represent 95% confidence intervals calculated using the subject-normalized method of Morrey (2008).

this difference reached the threshold of statistical significance ($t(2635.650) = 2.41$, $p = < .01$). A second contrast of these two conditions, this time collapsing over the retention interval, indicated a smaller mean difference (about 4% slower in the whole object study

condition than in the guided convergent retrieval condition) that also reached the level of statistical significance ($t(2645.2) = 2.106$, $p < .05$).

4.4 Discussion

The results of Experiment 3 confirm a counter-intuitive prediction of the PCR model – a method of study that provides less overall exposure, but where information unfolds sequentially, can produce better performance than a standard "free study" method. Objects studied using the guided convergent retrieval method, where object features were revealed one at time, produced faster and more accurate recall than when participants were allowed to study the whole object. This effect is predicted by the mechanism of the PCR model that explains better retention of information after retrieval than after restudy, convergent retrieval learning. Convergent retrieval learning (i.e., intra-item learning) is assumed to occur when the features representing a memory become active in a staged, sequential manner. Here, the same mnemonic benefits seen after a practice test were achieved by engineering the same process of staged, sequential activation to occur during the study phase, rather than relying on retrieval during a memory test to induce it.

While this experiment was inspired by a theory explaining what is learned from retrieval, and the experimental paradigm was successful at inducing retrieval-like mnemonic benefits, the pattern of recall accuracy here doesn't match what is typically seen in an experiment that contrasts retrieval practice with restudy practice. In typical "testing effect" experiments, a crossover interaction between retention interval and practice method is observed (e.g., see Figure 1, Figure 7 and Figure 12); restudy produces higher accuracy in the short term, while retrieval practice produces higher accuracy in the

long term. In Experiment 3 however, guided convergent retrieval practice (specifically designed to mimic learning from a test) produced better accuracy than whole object study at *both* retention intervals. The reason for this consistent advantage, rather than a temporally graded one, relates back to the bifurcated distribution model discussed in the introduction. In a typical retrieval practice experiment, accuracy on the practice test is less than perfect, meaning learning takes place for some items (the retrieved ones), but not for others (the unretrieved ones). Because restudy is free to strengthen all items in that condition, it takes time for the difference in memory strength to be revealed by forgetting. In this experiment, no such bifurcation takes place because guided convergent retrieval practice is applied equally to each object in that condition. Thus, without the performance-based bifurcation, accuracy is higher for all such items at both the immediate and delayed final test.

When introducing and explaining the PCR model, the features of the memory representations are abstract units, assumed to represent perceptual, semantic, phonological, or orthographic attributes and conjunctions of a to-be-remembered episode. In this experiment, a more concrete definition of the features needed for recall is used, operationalizing them as different parts of the studied objects. Thus, arguing that the results of Experiment 3 are accounted for by the intra-item learning mechanism that is proposed in the PCR model is to assume that visual objects are decomposed into representations of distinct parts, and that associations between these representations can be learned. This assumption is supported theoretically by hierarchical memory models (e.g., Cowell, Bussey, & Saksida, 2006) that assume processing of visual information proceeds along a continuum, starting with representations of simple visual features, to

feature conjunctions, to object identity, and all the way to associations of the visual information with episodic contexts. These models include representations of object parts that lie intermediate between simple features and identity, and hold that associations between these parts can be learned and used to retrieve other visual object knowledge at the same level in the hierarchy.

Supporting this account of visual memory, Sadil, Potter, Huber, & Cowell (2019) demonstrated a behavioral dissociation between knowledge about an object's visual details and knowledge of the object's identity. In this experiment, participants studied images of objects under Continuous Flash Suppression (CFS, a stimulus presentation technique which prevents visual information from reaching awareness) and studied the names of other objects under normal viewing conditions[13]. Later, they were tested on their memory for these objects in two ways. One test was a two-alternative matching test, in which participants saw two pairs of object parts, and had to judge which pair contained parts from the same object and which pair contained parts from two different objects. In the second test, participants were shown a single part from an object (which had either been studied under CFS, or whose name had been studied) and had to name the object the part came from (similar to the test administered here in Experiment 3). Participants were more accurate at discriminating the correctly paired parts from the incorrectly paired parts in the matching task when the object was studied under CFS as compared to when the object's name was studied, but were better able to report the object's identity from a single part after studying just the object's name than after studied the object itself under

[13] Studying the objects under CFS was necessary to allow specifically visual representations to be formed, but prevent the participant from consciously encoding the object's identity.

CFS. A state-trace analysis (Bamber, 1979) of accuracy in the matching task and the naming task using a hierarchical Bayesian model provided strong evidence that multiple processes were necessary to explain the results: one for the part-to-part associations and one for the part-to-identity associations. In other words, participants were able to learn which object parts went together independently of learning what the object was.

These findings support the idea that guided convergent retrieval practice helped participants learn associations between the object's visual features, which could be leveraged to pattern-complete the visual object when prompted with a single feature on the final test. Once the complete visual details of the object were retrieved, they could then retrieve a name for the well-known object. Such a retrieval process (i.e., retrieving visual details, then retrieving a name) may also explain the reason why recall accuracy was consistently different between the whole object study condition and the guided retrieval convergent retrieval condition, but recall latencies were not. If participants were better able to retrieve the complete visual details of an object, they would also be more likely to be able to recall a name for it (i.e., you can name the object only after you "see" it). However, neither condition gave participants experience retrieving the names of objects. If retrieving the visual details of the object was a relatively fast process (i.e., explained a relatively small proportion of the overall variability in naming latency) while retrieving a name for the object was a relatively slow process, (i.e., explained a relatively large proportion of the overall variability in naming latency), then variability in retrieval speed for the verbal label for the object could easily overshadow any difference in the speed of visual information retrieval, making it difficult to detect a reliable difference in the speed of recall. However, this explanation is admittedly post-hoc. The PCR model

itself predicts that the speed of retrieval should be increased when intra-item learning takes place, and observing this speed-up should not directly depend on the retention interval. While it is true that the statistical model indicated that there was no interaction between the two practice conditions and retention interval (allowing the conclusion that the relationship between the whole object study condition and the guided retrieval convergent retrieval condition were stable and different over time, as predicted by the PCR model), the strong similarity of the recall latencies from these conditions on the immediate final test was not an *a priori* prediction. This relationship should be investigated further with paradigms that do not require explicit naming, such as the two-alternative matching judgment used by Sadil et al. (2019), to avoid any confound from retrieving verbal labels.

There remains one caveat to the conclusions from this experiment, related to cue exposure at the onset of study for each item. In the guided retrieval convergent retrieval condition, the object feature that was initially displayed to begin the sequence of unfolding features was the same feature given as the retrieval cue on the test. This was done for principled reasons: the PCR model assumes directional associations are learned during convergent retrieval, and taking advantage of these associations requires the same "start pointing" be activated on both tests. Thus, the feature beginning the cascade that reveals the object is the "starting point" and intra-item learning should be evident when this feature is used to cue retrieval. However, this procedure also provided 250 ms. of exposure to the retrieval cue alone, which was not provided in the whole object study condition. Participants did not know that this feature would serve as the retrieval cue on the final test, as the experiment was not blocked into lists, and study trials for all items in

the experiment were completed before any test trials were administered. Still, the possibility exists that participants had more accurate memory in the guided retrieval convergent retrieval condition because of their brief exposure to the retrieval cue alone. Addressing this potential methodological issue is challenging, because it is difficult to strongly test the PCR model's predictions about this very situation (i.e., retrieval beginning from a specific point of partial activation) while keeping all presentation duration and formats the same. For example, one might address this problem of differential cue exposure by displaying the cue feature alone just before the object in the whole object study condition. However, this format (cue → whole object) may induce the very kind of pattern completion learning (i.e., going from part to whole) that is meant to be contrasted with study/restudy, and produce a null result where it is impossible to tell whether both conditions induced intra-item learning, or neither did. Future follow-up to this study should investigate accuracy in this task based on seeing just the cue alone, as well as the directionality assumption of the PCR model; if directionality of intra-item learning is not a hard requirement, then other features than the one that begins the unfolding sequence could be used as retrieval cues instead.

CHAPTER 5

GENERAL DISCUSSION

5.1 Summary of Findings

The Primary and Convergent Retrieval model of recall builds on the assumption made by most memory models that recall is a two-stage process. An initial stage (primary retrieval) uses context and any other cues provided 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. 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 following the temporal order in which features become active, the PCR model predicts that recalling an item 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. Thus, the PCR model provides a mechanistic explanation of the classic "testing effect".

Along with further specifying the recovery process, and nature of learning specific to recovery, the PCR model specifies the dynamic time course of recall. It predicts that recall latency is determined not only by the efficacy of retrieval cues, but by the unfolding of the convergent retrieval process. As a result of intra-item learning, it both more likely that an item will be recalled on future memory tests, and recalling the item should take less time (i.e., convergence occurs in fewer time steps). Previous research has confirmed this prediction, showing that retrieval practice produces faster recall latencies than restudy at both immediate and delayed final tests (Hopper & Huber, 2018; van den Broek et al., 2014). Critically, accuracy and latency changes dissociate on an immediate final test when comparing the effects of restudy versus a practice recall test, as predicted by the PCR model: restudy primarily increases memory accuracy

relative to baseline, whereas test practice primarily decreases retrieval latency. This supports the idea that retrieval practice produces qualitatively different learning than restudy practice, which takes the form of strengthened intra-item associations in the PCR model.

The three experiments reported here were designed to seek further evidence that the learning unique to retrieval is in fact strengthened feature-to-feature associations that are used during recovery. To do this, I tested behavioral predictions derived from the specific assumptions of the PCR model: when information is gradually activated, and when similar information is initially active during both practice and final test, recall accuracy should be increased and recall latency should be decreased. Experiment 1 tested the prediction that learning from testing applied to unretrieved and retrieved items alike, finding support for this prediction in the form of faster recall *failures* as well as faster recall successes after test practice. Experiment 2 demonstrated that learning from testing is specific to the cues used during recall practice, suggesting that associative information unique to the episodic conjunctions formed during study are a critical component of the features stored in memory traces. Experiment 3 shows that gradual unfolding of information produces better learning and retention than longer traditional study, confirming a counterintuitive prediction of the PCR model's intra-item learning mechanism. These results support the PCR model account of retrieval-based learning, and provide further information to guide future model developments.

5.2 Competing Accounts of Retrieval Practice Learning

The PCR model is far from the first explanation of learning from recall practice. In this final section, I discuss the differences between the PCR model and other theories, as well as their similarities.

5.2.1 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 – recalling requires recall processes, therefore utilizing recall processes during learning is optimal. Despite its intuitive and common sense appeal, transfer-appropriate processing remains 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 itself can be seen as a specific 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.

5.2.2 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 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 over the long term). 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 disuse, it is unclear which memory strength maps onto latency).

5.2.3 Dual Memory.

As mentioned in the discussion of Experiment 2, the Dual Memory theory holds that the difference between learning from restudy and learning from retrieval lies in which memory traces are strengthened/created. Restudy is assumed to strengthen the memory trace formed during initial study, while retrieval both strengthens the initial trace (on which practice test retrieval is based) *and* encodes a new, separate test memory trace. Thus, there is greater retention for retrieved items relative to restudy because there are two memory traces to rely of for those items, instead of just one. The Dual Memory theory shares some similarities with the theory of disuse, in that it holds that restudy and test have qualitative distinct effects on memory (rather retrieval just being a more effective form of restudy). However, it is different in that it holds there are distinct memory *traces* for study and test episodes, rather than there being two distinct memory

strength values affected by study and retrieval. The Dual Memory account directly conflicts with the PCR model, which assumes retrieval strengthens an existing memory trace. Memory models commonly assume repeated exposure to an item strengthens a single memory trace rather than encoding multiple traces, though there are successful exceptions to this rule (Hintzman, 1988). In general, it is difficult to distinguish multiple trace models from single trace models because they make such similar predictions (Atkinson & Shiffrin, 1965; Humphreys, Pike, Bain, & Tehan, 1989). In support of the idea that test practice encodes a new memory with qualitatively different structure than initial study, Rickard and Pan (2017) cite evidence showing that when keyword mediators presented during the learning of foreign vocabulary pairs are used in a lexical decision task given immediately after a translation practice trial, there is evidence of priming (i.e., faster lexical decision response times) when they are tested after a few translation practice trial but not after many trials. They argue this shows support for a shift from relying on the study memory (which includes the keyword mediator) to perform translation, to relying on the test memory (which enables translation via a direct cue-response association). Pan and Rickard also argue that changes to the recall latency distribution for retrieval of word triplets from a single power-law distribution to a mixture of power-law distributions reflects a change in reliance on study memory to reliance on test memory. The PCR model would explain this change as reflecting learning in the convergent retrieval learning process, and it is likely that a formal modeling approach would be required to distinguish these two accounts of the response time changes.

There may also be other qualitative ways to discriminate between these two accounts, for example, by examining the effects of retrieval practice on other memory effects, such as the list-strength effect in recognition. In recognition memory, a somewhat counter-intuitive finding has been that mixing together strongly encoded items with weakly encoded items on the same test list doesn't reduce recognition performance for the weak items (or enhance recognition performance for the strong items) relative to when the list is composed solely of strongly or weakly encoded test items. At the time, this finding of a so-called "null list strength effect" presented a fundamental challenge to existing memory models (Shiffrin, Ratcliff, & Clark, 1990). The REM model, developed in response to this challenging finding, accounted for the null list strength effect via a differentiation mechanism: strengthening an item in memory increased its dissimilarity to other words (such as the lures used on a recognition test), which prevented overall memory activation from becoming increasingly noisy (i.e., variable) with repeated study (Shiffrin & Steyvers, 1997). An important part of the REM model's differentiation mechanism was that learning from all strengthening operations (such as additional study time or study opportunities) were stored in the same memory trace. Without this single trace assumption, strengthening operations would function in a similar fashion as adding more and more items to the studied list, a manipulation which *does* consistently reduce recognition performance (i.e., a list-length effect)[14]. Though the following prediction will

[14] There is at least *some* flexibility in the REM model's ability to produce a null list-strength effect when storing new traces as opposed to strengthening existing traces. Shiffrin and Steyvers (1997) explored using study-phase retrieval success as a mechanism for deciding whether to update an existing trace or store a new on in their REM.3 simulations. This implementation stored new traces infrequently (i.e., retrieval of an existing trace to update was usually successful), and was still able to produce a null list strength effect consistent with the observed data.

depend on the precise nature of the cue-response association that Pan and Rickard have suggested, the general proposal that a test trial encodes a new memory trace would predict that interspersing retrieval practice between a studying a list of mixed strong and weak items and a final recognition test may be able to induce a positive list-strength effect. The PCR model predicts that retrieval practice would not affect the list-strength effect in recognition differently than restudy, because it is assumed the intra-item associations are not used to support recognition decisions, and thus strengthening them would not impact recognition performance.

5.2.4 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).

5.2.5 Episodic Context

Another recently proposed explanation of learning from retrieval practice appeals to the effects of updating of stored contextual representations during testing. According to the episodic context account, tests produce better long term retention than restudy because

recall causes the context representations associated with the retrieved information to be updated with features of the test context, adding to the features of the study context, and pre-experimental context (Karpicke, Lehman, & Aue, 2014; Lehman, Smith, & Karpicke, 2014). This updating benefits retrieval in several ways. Because retrieved items are updated to reflect the context of the practice test, they are stored with a context that is more similar to the context of the final test, facilitating retrieval. This updating also increases the diversity of the contextual features associated with tested item, increasing the chances that a future contextual state (e.g., on a delayed final test) will overlap with it and activate the item again. If this updating does not occur with restudy, the context associated with studied items with be both less diverse and more similar to the past than the present, causing weaker retention following restudy than retrieval practice. This account has the advantage of being well-integrated with established theories of memory retrieval, building upon existing retrieved context models which successfully explain organizational patterns of free recall behavior (Howard & Kahana, 2002; Lehman & Malmberg, 2013; Polyn, Norman, & Kahana, 2009). None of its assumptions about contextual associations, contextual reinstatement, or contextual updating are particularly controversial or new, and in general, should be appreciated for providing a coherent explanation for how many different memory processes that have been previously proposed could be combined to explain retrieval-based learning. However, I perceive a potentially serious complication for this theory as it has been put forth so far: the computational models from which its assumptions about contextual updating are derived assume that contextual updating occurs during study, while this account holds that contextual updating explains the unique *difference* between restudy and testing.

Furthermore, the assumptions about contextual updating during study are central mechanisms within these families of models. For example, the Temporal Context Model (TCM; Howard & Kahana, 2002) relies upon the updating of the current context with the pre-experimental context features associated with each item to account for the asymmetry in transitions between the serial positions of recalled items (i.e., forward transitions are more likely than backwards transitions). Still, this problem is not necessarily fatal for the episodic context account. The updating assumptions could be relaxed to allow the degree of updating during restudy and test to simply be weighted differently, or by assuming that the onset of the practice test induces a sharp context shift, meaning that contextual diversity is improved for recalled items relative to restudied items. It remains to be seen whether relaxing these assumptions will cause the model to sacrifice the ability to explain prior organization effect (like the transition probabilities) in order to accommodate the testing effect.

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 due to a having a smaller 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 (i.e., convergent retrieval in PCR) rather than the sampling process (i.e., primary retrieval in PCR). Hopper and Huber (2018) compared these accounts 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, despite a shared context between practice and final tests. 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. It is also difficult to explain why restudy (an operation which should increase the size of a memory search set under nearly any theoretical perspective) does not slow recall latencies after restudy despite increasing accuracy (Hopper & Huber, 2018). Comparisons of the PCR model and the episodic context account will require further investigation of free recall latencies (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.

TABLES

Table 1: AIC and BIC goodness of fit statistics for the six LBA models applied to each participant. The model with the lowest BIC/AIC is the most preferred model, and is denoted for each participant with an asterisk.

Subject	Statistic	*v* Free	*A* Free	*b* Free	T_0 Free	*v* & *A* Free	*v* & *b* Free
1	AIC	1611.18	1718.87	1652.6	1706.9	1610.10	*1565.42
	BIC	*1679.11	1795.30	1729.05	1787.5	1724.74	1680.06
2	AIC	*1592.39	1653.47	1638.18	1644.4	1594.49	1605.52
	BIC	*1655.09	1724.53	1709.24	1719.6	1698.99	1710.02
3	AIC	1646.47	1661.62	1614.60	1660.7	1655.94	*1613.71
	BIC	1701.64	1725.28	*1678.27	1728.5	1745.06	1702.84
4	AIC	2042.22	2119.18	2060.09	2101.1	2037.25	*2014.05
	BIC	*2116.66	2201.89	2142.80	2188.0	2165.46	2142.26
5	AIC	*1890.13	1991.22	1958.48	1980.3	1901.27	1917.19
	BIC	*1959.21	2068.93	2036.19	2062.3	2017.83	2033.75
6	AIC	1969.69	2075.32	1901.01	2046.7	1954.51	*1830.97
	BIC	2039.59	2153.44	1979.13	2128.9	2073.74	*1950.21
7	AIC	1890.10	1996.22	1923.43	1982.4	1888.92	*1850.15
	BIC	*1962.05	2076.64	2003.85	2067.0	2011.66	1972.89
8	AIC	1982.94	2018.70	1907.33	2020.8	1977.62	*1885.32
	BIC	2050.08	2094.24	*1982.87	2100.5	2090.92	1998.63
9	AIC	1591.67	1584.60	1542.02	1586.0	1561.47	*1511.76
	BIC	1656.08	1657.06	*1614.48	1662.5	1670.17	1620.46
10	AIC	2026.30	2054.72	2090.41	2043.4	*2015.20	2062.61
	BIC	*2103.92	2140.95	2176.64	2134.0	2148.86	2196.28
Total	**AIC**	18243.09	18873.92	18288.17	18773.0	18196.76	*17856.72
	BIC	*18923.44	19638.26	19052.51	19579.4	19347.44	19007.40

Sub.	Drift Rate	Immediate Test				Delayed Test			
		None	Restudy	Test Practice, Incorrect	Test Practice, Correct	None	Restudy	Test Practice, Incorrect	Test Practice, Correct
1	v_0	0.670	0.393	1.514	-	1.288	1.169	1.593	0.885
	v_1	1.326	1.812	-	2.161	0.399	0.636	-	1.216
2	v_0	0.623	0.162	1.625	-	1.287	1.271	1.178	1.322
	v_1	1.010	1.593	-0.190	2.070	-	-	-	-0.309
3	v_0	-	-	-	-	1.083	1.001	1.110	1.039
	v_1	1.830	2.104	-	1.920	-0.315	0.341	-	0.142
4	v_0	0.510	0.206	0.792	-0.089	0.859	0.624	0.780	0.489
	v_1	0.672	1.207	-1.558	1.999	-1.715	-1.854	-	-0.066
5	v_0	1.009	0.805	2.034	-0.237	1.098	1.229	1.568	0.837
	v_1	1.058	1.286	-	2.278	-	-0.752	-	-0.361
6	v_0	0.351	0.285	1.602	-	0.627	0.510	0.726	0.456
	v_1	0.801	1.559	-0.027	2.433	-1.489	-0.266	-	-0.029
7	v_0	-0.029	-0.907	1.552	-0.268	1.289	0.883	1.500	0.524
	v_1	1.102	1.548	-	1.989	-1.113	-0.157	-	0.389
8	v_0	0.411	0.440	1.681	-	0.712	0.684	0.792	0.693
	v_1	1.306	1.571	-	1.988	-0.758	-0.520	-	0.306
9	v_0	-0.796	-	0.373	-	0.319	0.289	0.246	-0.021
	v_1	1.484	1.699	0.917	1.946	-1.124	-0.152	-	0.380
10	v_0	0.704	0.678	1.295	0.398	1.322	1.275	1.282	1.210
	v_1	0.875	1.102	0.202	1.466	-0.398	-0.059	-	-0.128

Table 2: Best fitting starting point variability and non-decision time parameters for the "v Free" model. The 0 and 1 subscripts refer to the "Can't Recall" (incorrect) and "Recall" (correct) accumulators, respectively. The b and s parameters were set to constant values of 4 and .5, respectively.

Subject	A_0	A_1	T_0
1	2.48	2.41	0.00
2	3.80	3.36	0.91
3	0.00	2.96	0.29
4	3.46	3.79	0.48
5	3.70	3.33	0.68
6	0.00	3.67	0.71
7	3.77	3.66	0.75
8	0.00	2.89	0.00
9	2.58	2.77	0.00
10	3.30	2.71	0.01

Table 3: Prior distributions for recall accuracy regression coefficients on the log-odds scale. Priors are appropriate for a regression using sum-coded dummy regressors with the delayed final test and no practice conditions set as the reference levels.

Parameter	Prior Distribution	Source
Intercept (Grand Mean)	$\sim N\begin{pmatrix}\mu = 0.268\\ \sigma = 0.393\end{pmatrix}$	Hopper & Huber, 2018
Retention Interval (Immediate Condition)	$\sim N\begin{pmatrix}\mu = 0.837\\ \sigma = 0.393\end{pmatrix}$	Hopper & Huber, 2018
Episodic Cue Restudy	$\sim N\begin{pmatrix}\mu = 0.635\\ \sigma = 0.156\end{pmatrix}$	Hopper & Huber, 2018
Semantic Cue Restudy	$\sim N\begin{pmatrix}\mu = 0.252\\ \sigma = 0.126\end{pmatrix}$	Dissertation Pilot Study
Episodic Cue Test	$\sim N\begin{pmatrix}\mu = -0.008\\ \sigma = 0.141\end{pmatrix}$	Hopper & Huber, 2018
Semantic Cue Test	$\sim N\begin{pmatrix}\mu = 0.142\\ \sigma = 0.122\end{pmatrix}$	Dissertation Pilot Study
Episodic Cue Restudy × Immediate Final Test	$\sim N\begin{pmatrix}\mu = 0.406\\ \sigma = 0.126\end{pmatrix}$	Hopper & Huber, 2018
Semantic Cue Restudy × Immediate Final Test	$\sim N\begin{pmatrix}\mu = 0\\ \sigma = 2.5\end{pmatrix}$	Weak Prior (rstan default)
Episodic Cue Test × Immediate Final Test	$\sim N\begin{pmatrix}\mu = -0.354\\ \sigma = 0.126\end{pmatrix}$	Hopper & Huber, 2018
Semantic Cue Test × Immediate Final Test	$\sim N\begin{pmatrix}\mu = 0\\ \sigma = 2.5\end{pmatrix}$	Weak Prior (rstan default)

Table 4: Posterior probabilities of recall for each condition in Experiment 2.

Final Test	Practice Type	Median Proportion Correct	95% HDI Bounds
Immediate	Episodic Cue Restudy	0.796	[0.742, 0.848]
	Semantic Cue Restudy	0.631	[0.555, 0.703]
	Semantic Cue Test	0.570	[0.493, 0.647]
	Episodic Cue Test	0.558	[0.478, 0.637]
	No Practice	0.473	[0.397, 0.557]
Delay	Episodic Cue Restudy	0.379	[0.298 0.457]
	Semantic Cue Restudy	0.295	[0.229, 0.367]
	Semantic Cue Test	0.261	[0.198, 0.327]
	Episodic Cue Test	0.392	[0.314, 0.474]
	No Practice	0.213	[0.157, 0.273]

Table 5: Contrasts of recall accuracy between practice conditions at each retention interval in Experiment 2.

Final Test	Contrast	Median Log-odds Difference	95% HDI Bounds
Immediate	Semantic Cue Restudy – Episodic Cue Restudy	-0.824	[-1.003, -0.644]
	Semantic Cue Test – Semantic Cue Restudy	-0.255	[-0.429, -0.088]
	Episodic Cue Test – Semantic Cue Test	-0.047	[-0.222, 0.111]
	Episodic Cue Test – Episodic Cue Restudy	-1.124	[-1.306, -0.942]
	No Practice – Episodic Cue Test	-0.341	[-0.515, -0.175]
	No Practice – Semantic Cue Test	-0.386	[-0.547, -0.207]
	No Practice – Semantic Cue Restudy	-0.642	[-0.825, -0.472]
	No Practice – Episodic Cue Restudy	-1.467	[-1.658, -1.287]
Delay	Semantic Cue Restudy –Episodic Cue Restudy	-0.376	[-0.556, -0.200]
	Semantic Cue Test – Semantic Cue Restudy	-0.171	[-0.354, 0.003]
	Episodic Cue Test – Semantic Cue Test	0.604	[0.417, 0.780]
	Episodic Cue Test – Episodic Cue Restudy	0.056	[-0.118, 0.239]
	No Practice – Episodic Cue Test	-0.868	[-1.071, -0.675]
	No Practice – Semantic Cue Test	-0.263	[-0.473, -0.069]
	No Practice – Semantic Cue Restudy	-0.436	[-0.637, -0.245]
	No Practice – Episodic Cue Restudy	-0.813	[-1.020, -0.623]

Table 6: Interaction contrasts of recall accuracy in Experiment 2. SC-TP = Semantic Cue Test Practice condition, SC-RS = Semantic Cue Restudy condition, EC-TP = Episodic Cue Test Practice condition, EC-RS = Episodic Cue Restudy condition. The "Imm" and "Del" subscripts refer to the immediate and delayed final test conditions, respectively.

Interaction Contrast	Median Log-odds Difference	95% HDI Bounds
$(\text{SC-TP}_{\text{Imm}} - \text{SC-RS}_{\text{Imm}}) - (\text{SC-TP}_{\text{Del}} - \text{SC-RS}_{\text{Del}})$	-0.082	[-0.327, 0.174]
$(\text{EC-TP}_{\text{Imm}} - \text{EC-RS}_{\text{Imm}}) - (\text{EC-TP}_{\text{Del}} - \text{EC-RS}_{\text{Del}})$	-1.181	[-1.439, -0.933]
$(\text{EC-TP}_{\text{Imm}} - \text{SC-TP}_{\text{Imm}}) - (\text{EC-TP}_{\text{Del}} - \text{SC-TP}_{\text{Del}})$	-0.651	[-0.880, -0.391]

Table 7: Prior distributions for recall latency regression coefficients on the natural log scale. Priors are appropriate for a regression using sum-coded dummy regressors with the delayed final test and no practice conditions set as the reference levels.

Parameter	Prior Distribution	Source
Intercept (Grand Mean)	$\sim N\begin{pmatrix}\mu = 0.632\\ \sigma = 0.066\end{pmatrix}$	Hopper & Huber, 2018
Retention Interval (Immediate Condition)	$\sim N\begin{pmatrix}\mu = -0.213\\ \sigma = 0.066\end{pmatrix}$	Hopper & Huber, 2018
Episodic Cue Restudy	$\sim N\begin{pmatrix}\mu = -0.015\\ \sigma = 0.033\end{pmatrix}$	Hopper & Huber, 2018
Semantic Cue Restudy	$\sim N\begin{pmatrix}\mu = -0.046\\ \sigma = 0.036\end{pmatrix}$	Dissertation Pilot Study
Episodic Cue Test	$\sim N\begin{pmatrix}\mu = -0.171\\ \sigma = 0.033\end{pmatrix}$	Hopper & Huber, 2018
Semantic Cue Test	$\sim N\begin{pmatrix}\mu = -0.046\\ \sigma = 0.036\end{pmatrix}$	Dissertation Pilot Study
Episodic Cue Restudy × Immediate Final Test	$\sim N\begin{pmatrix}\mu = 0.006\\ \sigma = 0.033\end{pmatrix}$	Hopper & Huber, 2018
Semantic Cue Restudy × Immediate Final Test	$\sim N\begin{pmatrix}\mu = 0\\ \sigma = 2.5\end{pmatrix}$	Weak Prior (rstan default)
Episodic Cue Test × Immediate Final Test	$\sim N\begin{pmatrix}\mu = -0.032\\ \sigma = 0.033\end{pmatrix}$	Hopper & Huber, 2018
Semantic Cue Test × Immediate Final Test	$\sim N\begin{pmatrix}\mu = 0\\ \sigma = 2.5\end{pmatrix}$	Weak Prior (rstan default)

Table 8: Posterior recall latency for each condition in Experiment 2.

Final Test	Practice Type	Median log-scale Recall Latency	95% HDI Bounds
Immediate	Episodic Cue Restudy	0.507	[0.431, 0.576]
	Semantic Cue Restudy	0.579	[0.505, 0.654]
	Semantic Cue Test	0.617	[0.541, 0.692]
	Episodic Cue Test	0.413	[0.335, 0.485]
	No Practice	0.718	[0.642, 0.796]
Delay	Episodic Cue Restudy	0.775	[0.692, 0.851]
	Semantic Cue Restudy	0.861	[0.783, 0.945]
	Semantic Cue Test	0.817	[0.738, 0.902]
	Episodic Cue Test	0.673	[0.599, 0.755]
	No Practice	0.976	[0.893, 1.067]

Table 9: Contrasts of recall latency between practice conditions at each retention interval in Experiment 2.

Final Test	Contrast	Median Log-scale Difference	95% HDI Bounds
Immediate	Semantic Cue Restudy — Episodic Cue Restudy	0.072	[0.026, 0.114]
	Semantic Cue Test — Semantic Cue Restudy	0.037	[-0.009, 0.083]
	Episodic Cue Test — Semantic Cue Test	-0.204	[-0.250, -0.159]
	Episodic Cue Test — Episodic Cue Restudy	-0.094	[-0.137, -0.049]
	No Practice — Episodic Cue Test	0.305	[0.253, 0.354]
	No Practice — Semantic Cue Test	0.102	[0.050, 0.152]
	No Practice — Semantic Cue Restudy	0.139	[0.089, 0.190]
	No Practice — Episodic Cue Restudy	0.211	[0.165, 0.261]
Delay	Semantic Cue Restudy — Episodic Cue Restudy	0.087	[0.030, 0.143]
	Semantic Cue Test — Semantic Cue Restudy	-0.044	[-0.108, 0.018]
	Episodic Cue Test — Semantic Cue Test	-0.144	[-0.204, -0.086]
	Episodic Cue Test — Episodic Cue Restudy	-0.102	[-0.158, -0.049]
	No Practice — Episodic Cue Test	0.303	[0.239, 0.365]
	No Practice — Semantic Cue Test	0.159	[0.091, 0.230]
	No Practice — Semantic Cue Restudy	0.115	[0.044, 0.179]
	No Practice — Episodic Cue Restudy	0.201	[0.138, 0.266]

Table 10: Interaction Contrasts of recall latency in Experiment 2. SC-TP = Semantic Cue Test Practice condition, SC-RS = Semantic Cue Restudy condition, EC-TP = Episodic Cue Test Practice condition, EC-RS = Episodic Cue Restudy condition. The "Imm" and "Del" subscripts refer to the immediate and delayed final test conditions, respectively.

Interaction Contrast	Median Log-scale Difference	95% HDI Bounds
$(\text{EC-RS}_{\text{Imm}} - \text{EC-TP}_{\text{Imm}}) - (\text{EC-RS}_{\text{Del}} - \text{EC-TP}_{\text{Del}})$	-0.007	[-0.077, 0.060]
$(\text{SC-RS}_{\text{Imm}} - \text{SC-TP}_{\text{Imm}}) - (\text{SC-RS}_{\text{Del}} - \text{SC-TP}_{\text{Del}})$	-0.082	[-0.161, -0.001]
$(\text{SC-TP}_{\text{Imm}} - \text{EC-TP}_{\text{Imm}}) - (\text{SC-TP}_{\text{Del}} - \text{EC-TP}_{\text{Del}})$	0.060	[-0.018, 0.133]

Table 11: Interaction contrasts of naming accuracy in Experiment 3. N, S, and GCR condition abbreviations stand for No Practice, Whole Object Study, and Guided Convergent Retrieval, respectively. The "Imm" and "Del" subscripts refer to the immediate and delayed test conditions, respectively. *P*-values are adjusted using the Holm-Bonferroni method for 3 tests.

Contrast	Log-Odds Difference	Standard Error	Z	*P* value
$(N_{Imm} - S_{Imm}) - (N_{Del} - S_{Del})$	-1.076	0.192	-5.602	< .0001
$(N_{Imm} - GCR_{Imm}) - (N_{Del} - T_{Del})$	-1.026	0.191	-5.367	< .0001
$(S_{Imm} - GCR_{Imm}) - (S_{Del} - GCR_{Del})$	0.050	0.137	0.362	0.7171

Table 12: Interaction contrasts of naming latency in Experiment 3. N, S, and GCR condition abbreviations stand for No Practice, Whole Object Study, and Guided Convergent Retrieval, respectively. The "Imm" and "Del" subscripts refer to the immediate and delayed test conditions, respectively. *P*-values are adjusted using the Holm-Bonferroni method for 3 tests.

Contrast	Log-Scale Difference	Standard Error	*t* statistic	*P* value
$(N_{Del} - S_{Del}) - (N_{Imm} - S_{Imm})$	-0.368	0.081	$t(2633.8) = -4.55$	< .0001
$(N_{Del} - T_{Del}) - (N_{Imm} - GCR_{Imm})$	-0.300	0.080	$t(2631.7) = -3.74$	< .001
$(S_{Del} - GCR_{Del}) - (S_{Imm} - GCR_{Imm})$	0.068	0.043	$t(2638.4) = 1.58$	0.113

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Atkinson, R. C., & Shiffrin, R. M. (1965). *Mathematical Models for Memory and Learning* (Technical Report No. 79). Retrieved from Institute for Mathematical Studies in the Social Sciences website: http://cogs.indiana.edu/FestschriftForRichShiffrin/pubs/1965%20Mathematical%20Models%20for%20Memory%20and%20Learning.%20Shiffrin,%20Atkinson.pdf

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

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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. (1992-97939-014).

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

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

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

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., 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

Cowell, R. A., Bussey, T. J., & Saksida, L. M. (2006). Why Does Brain Damage Impair Memory? A Connectionist Model of Object Recognition Memory in Perirhinal Cortex. *Journal of Neuroscience*, *26*(47), 12186–12197. https://doi.org/10.1523/JNEUROSCI.2818-06.2006

Cox, G. E., & Shiffrin, R. M. (2017). A dynamic approach to recognition memory. *Psychological Review*, *124*(6), 795–860. https://doi.org/10.1037/rev0000076

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language*, *64*(4), 316–326.

Criss, A. H., & Shiffrin, R. M. (2004). Pairs do not suffer interference from other types of pairs or single items in associative recognition. *Memory & Cognition*, *32*(8), 1284–1297. https://doi.org/10.3758/BF03206319

Criss, A. H., & Shiffrin, R. M. (2005). List Discrimination in Associative Recognition and Implications for Representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1199–1212. https://doi.org/10.1037/0278-7393.31.6.1199

Diller, D. E., Nobel, P. A., & Shiffrin, R. M. (2001). An ARC–REM model for accuracy and response time in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(2), 414–435. https://doi.org/10.1037/0278-7393.27.2.414

Donkin, C., Brown, S. D., & Heathcote, A. (2009). The overconstraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review*, *16*(6), 1129–1135. https://doi.org/10.3758/PBR.16.6.1129

Donkin, C., Brown, S., & Heathcote, A. (2011). Drawing conclusions from choice response time models: A tutorial using the linear ballistic accumulator. *Journal of Mathematical Psychology*, *55*(2), 140–151. https://doi.org/10.1016/j.jmp.2010.10.001

Donkin, C., Brown, S., Heathcote, A., & Wagenmakers, E.-J. (2011). Diffusion versus linear ballistic accumulation: Different models but the same conclusions about psychological processes? *Psychonomic Bulletin & Review*, *18*(1), 61–69. https://doi.org/10.3758/s13423-010-0022-4

Dougherty, M. R., Harbison, J. I., & Davelaar, E. J. (2014). Optional Stopping and the Termination of Memory Retrieval. *Current Directions in Psychological Science*, *23*(5), 332–337. https://doi.org/10.1177/0963721414540170

Eich, J. M. (1985). Levels of Processing, Encoding Specificity, Elaboration, and CHARM. *Psychological Review*, *92*(1), 38. https://doi.org/10.1037/0033-295X.92.1.1

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., Cramon, D. Y. von, Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences*, *105*(45), 17538–17542. https://doi.org/10.1073/pnas.0805903105

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

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Oxford, England: John Wiley.

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

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528–551. https://doi.org/10.1037/0033-295X.95.4.528

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

Hopper, W. J., & Huber, D. E. (2018). Learning to recall: Examining recall latencies to test an intra-item learning theory of testing effects. *Journal of Memory and Language*, *102*, 1–15. https://doi.org/10.1016/j.jml.2018.04.005

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

Humphreys, M. S., Pike, R., Bain, J. D., & Tehan, G. (1989). Global matching: A comparison of the SAM, Minerva II, Matrix, and TODAM models. *Journal of Mathematical Psychology*, *33*(1), 36–67. https://doi.org/10.1016/0022-2496(89)90003-5

Karpicke, J. D. (2017). Retrieval-Based Learning: A Decade of Progress. In J. T. Wixted (Ed.), *Cognitive psychology of memory, Vol. 2 of Learning and memory: A comprehensive reference*. https://doi.org/10.1016/B978-0-12-809324-5.21055-9

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). 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

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*(3), 983–997.

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

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 283–294. https://doi.org/10.1037/a0037850

Kuo, T.-M., & Hirshman, E. (1996). Investigations of the testing effect. *The American Journal of Psychology*, *109*(3), 451–464. https://doi.org/10.2307/1423016

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

Lenth, R. (2018). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. Retrieved from https://CRAN.R-project.org/package=emmeans

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

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

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide, 2nd ed.* Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Malmberg, K. J. (2008). Investigating Metacognitive Control in a Global Memory Framework. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of Metamemory and Memory*. https://doi.org/10.4324/9780203805503.ch14

Malmberg, K. J., & Shiffrin, R. M. (2005). The "One-Shot" Hypothesis for Context Storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 322–336. https://doi.org/10.1037/0278-7393.31.2.322

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in Middle-School Science: Successful Transfer Performance on Classroom Exams: Quizzing and successful transfer. *Applied Cognitive Psychology*, *27*(3), 360–372. https://doi.org/10.1002/acp.2914

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. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*(6), 609–626. https://doi.org/10.1037/0033-295X.89.6.609

Murdock, B. B. (1993). TODAM2: A model for the storage and retrieval of item, associative, and serial-order information. *Psychological Review*, *100*(2), 183–203. https://doi.org/10.1037/0033-295X.100.2.183

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407. https://doi.org/10.3758/BF03195588

Nelson, T. O., & Dunlosky, J. (1991). When People's Judgments of Learning (JOLs) are Extremely Accurate at Predicting Subsequent Recall: The "Delayed-JOL Effect." *Psychological Science*, *2*(4), 267–271. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

Nobel, P. A., & Shiffrin, R. M. (2001). Retrieval processes in recognition and cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(2), 384–413. https://doi.org/10.1037/0278-7393.27.2.384

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

Osth, A. F., Bora, B., Dennis, S., & Heathcote, A. (2017). Diffusion vs. linear ballistic accumulation: Different models, different conclusions about the slope of the zROC in recognition memory. *Journal of Memory and Language*, *96*, 36–61. https://doi.org/10.1016/j.jml.2017.04.003

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

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.

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Retrieved from https://www.R-project.org/

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

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science*, *9*(5), 347–356. https://doi.org/10.1111/1467-9280.00067

Ratcliff, R., & Smith, P. L. (2004). A Comparison of Sequential Sampling Models for Two-Choice Reaction Time. *Psychological Review*, *111*(2), 333–367. https://doi.org/10.1037/0033-295X.111.2.333

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review*, *116*(1), 59–83. https://doi.org/10.1037/a0014086

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language*, *50*(4), 408–424. https://doi.org/10.1016/j.jml.2003.11.002

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*(3), 438–481.

Rickard, T. C., & Pan, S. C. (2017). A dual memory theory of the testing effect. *Psychonomic Bulletin & Review*, 1–23. https://doi.org/10.3758/s13423-017-1298-4

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., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

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

Sadil, P., Potter, K. W., Huber, D. E., & Cowell, R. A. (2019). Connecting the Dots Without Top-Down Knowledge: Evidence for Rapidly-Learned Low-Level Associations That Are Independent of Object Identity. *Journal of Experimental Psychology: General*, *148*(6), 1058–1070. https://doi.org/10.1037/xge0000607

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(2), 179–195. https://doi.org/10.1037/0278-7393.16.2.179

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166. https://doi.org/10.3758/BF03209391

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2018). *afex: Analysis of Factorial Experiments*. Retrieved from https://CRAN.R-project.org/package=afex

Singmann, H., Brown, S., Gretton, M., & Heathcote, A. (2017). *rtdists: Response Time Distributions*. Retrieved from https://CRAN.R-project.org/package=rtdists

Stan Development Team. (2016). *rstanarm: Bayesian applied regression modeling via Stan.* Retrieved from http://mc-stan.org/

Starns, J. J. (2014). Using response time modeling to distinguish memory and decision processes in recognition and source tasks. *Memory & Cognition*, *42*(8), 1357–1372. https://doi.org/10.3758/s13421-014-0432-z

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

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion Models in Experimental Psychology. *Experimental Psychology*, *60*(6), 385–402. https://doi.org/10.1027/1618-3169/a000218

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, J. H., & Criss, A. H. (2017). The list strength effect in cued recall. *Journal of Memory and Language*, *95*, 78–88. https://doi.org/10.1016/j.jml.2017.01.006