Memory & Cognition (2019) 47:816–841
https://doi.org/10.3758/s13421-019-00903-x

# Testing the primary and convergent retrieval model of recall: Recall practice produces faster recall success but also faster recall failure

William J. Hopper[1] · David E. Huber[1]

Published online: 8 February 2019

## Abstract

The primary and convergent retrieval (PCR) model assumes that the act of successful recall not only boosts associations between the item and retrieval cues but additionally strengthens associations within the item (i.e., between the features of an item), speeding the rate of information retrieval from memory. The latter effect is termed intra-item learning and is a unique benefit of recall practice (i.e., the "testing effect"). Prior work confirmed the prediction that recall practice produces faster subsequent recall than restudy practice even if accuracy is higher following restudy. The current study replicated this result, but also examined the downside of recall practice: that after a failure to recall during practice, participants will be faster in their failure to recall on a subsequent recall test. This prediction was confirmed in a multisession cued recall experiment that collected accuracy and recall latency measurements for no practice, recall practice, or restudy, with an immediate or delayed final test. The linear ballistic accumulator model was fit to latency distributions, and model comparison determined that these effects reflect differences in drift rates, as predicted by the PCR model.

**Keywords** Episodic memory · Cued recall · Retrieval practice · Cognitive modeling

Atkinson and Shiffrin (1968) proposed the modal model of memory, with separate storage "modes" differing in terms of information content, capacity, and retention duration. In addition, Atkinson and Shiffrin outlined the manner in which memories move through these modes, entering long-term storage in the form of separate memory traces for each episode, and subsequently, the manner in which information is retrieved from long-term storage. The current study concerns a specific aspect of this retrieval process—the manner in which an initial partial retrieval of a memory might or might not lead to full retrieval. A classic example of this is the tip-of-the-tongue phenomenon (Brown & McNeill, 1966), in which an individual might fail to recall something, but know that they know the desired information—for instance, as realized by their ability to recognize the answer. Of this phenomenon, Atkinson and Shiffrin wrote that "a simple trace model can probably not handle these results. A class of models for the trace which can explain the tip-of-the-tongue phenomenon are the multiple-copy models suggested by Atkinson and Shiffrin (1965)." (p. 105).

Atkinson and Shiffrin (1965) is a lesser known technical report, outlining much of what would appear in the subsequent 1968 paper, but this technical report went into greater detail regarding the mathematics of the proposed memory models. The single copy models that served as the basis for the 1968 paper are presented, but in addition, this technical report considered multiple-copy models in which different study episodes of the same item are stored in separate memory traces. First, it was demonstrated that the multiple trace model produces similar learning effects as compared with the strength-based single copy model. Second, a specific "information model" implementation of the multiple trace assumption was considered in which the copies are bits of information concerning an item. Thus, the longer that an item is studied (or the more times that it is encountered), the higher the proportion of stored features for that item. Of this information model, Atkinson and Shiffrin wrote, "On any one search this information may be insufficient to actually report the correct answer with assurance. On the other hand the idea of a small portion of information being available gives a natural explanation for the difference between recall and recognition measures of retention" (p. 80). This information model was not further developed, and the subsequent

✉ William J. Hopper
whopper@psych.umass.edu

[1] Department of Psychological and Brain Sciences, University of Massachusetts, 441 Tobin Hall, 135 Hicks Way, Amherst, MA 01003, USA

SAM (Gillund & Shiffrin, 1984; Raaijmakers & Shiffrin, 1981) and REM models (Malmberg & Shiffrin, 2005; Shiffrin & Steyvers, 1997), as applied to recall data, assume that the "recovery process" is based on the single-copy memory strength that also underlies recognition. In other words, the subsequent models assume that the main qualitative difference between recognition and recall arises from the sampling of traces in the memory search process.

The primary and convergent retrieval (PCR) model proposed by Hopper and Huber (2018) assumes a qualitatively different kind of associative information that is unique to recall, predicting differences between recall and recognition above and beyond the sampling process that only applies to recall. In other words, not only are recognition and recall different processes, as is assumed by all memory models, but the PCR model assumes they operate on different kinds of memory storage and are supported by different learning processes. The representation assumed in the PCR model is similar to Atkinson and Shiffrin's (1965) "information model," where memory traces are a pattern of features within a larger interconnected network. In the PCR model, the recovery process (i.e., convergent retrieval) that occurs during recall unfolds dynamically, with already retrieved item features serving as the cues for the retrieval of additional item features. Successful convergent retrieval not only enables recall of the item but additionally supports new learning for the associations between the features of an item. Thus, the PCR model proposes that the act of recalling something induces a kind of learning different from the learning that might occur when studying an item in a passive fashion. Furthermore, this convergent retrieval learning is predicted to affect the speed with which items are recalled from memory. In light of this proposal, the PCR model is well positioned to explain the benefits of retrieval practice, as briefly reviewed next.

## Retrieval practice

In preparation for an upcoming exam, which form of practice is most effective? Beyond rote rehearsal, a wide variety of techniques can be employed, including spaced learning, imagery, and "survival processing" (e.g., Nairne, Pandeirada, & Thompson, 2008). However, for long-term retention (e.g., preparing for an exam a week in advance, rather than 1 hour in advance), there is considerable evidence that the act of taking a practice test is particularly effective (for a review of early research into retrieval-based learning, see Roediger & Karpicke, 2006a, and see Karpicke, 2017, for a review of more recent findings), particularly if the practice test involves recall of the material (Rowland, 2014). Retrieval practice experiments typically have three phases: an initial acquisition phase, a practice phase, and a final test phase. The acquisition phase presents novel material for study (e.g., lists of unrelated word pairs, or text passages). In the practice phase, items from the acquisition phase are reviewed in different ways; typically, some items are restudied while a practice test is given for other items (e.g., recalling the missing word from a pair). Lastly, a memory test is administered for all items in the final test phase and the key comparisons concern final test accuracy following different kinds of practice. The benefits of test practice become apparent when there is a substantial retention interval (e.g., at least 24 hours) between the practice and final test; after a delay, memory retrieval is more accurate for items that received a practice test, as compared with restudied items, and this is true even in the absence of any feedback during the practice test (Carpenter & DeLosh, 2006; Carpenter, Pashler, Wixted, & Vul, 2008; Kuo & Hirshman, 1996; Roediger & Karpicke, 2006b; Toppino & Cohen, 2009; Wheeler, Ewers, & Buonanno, 2003).

The finding that recall practice is better than restudy after a delay might simply indicate that recall practice produces stronger memories. However, the effect of recall practice appears to be more complicated than a simple strengthening of all practiced items. For instance, if the final test occurs immediately after a practice phase that does not include feedback, the opposite pattern is often observed, with higher accuracy following restudy as compared with following test practice. Thus, there is a crossover interaction with delay, suggesting different forgetting rates following recall practice as compared with restudy. However, there is not much opportunity to learn from the failure to recall an item in the absence of feedback (although the current experiment demonstrates that something is indeed learned from recall failure—more specifically, this produces more rapid failure on the final test). Because not all items are recalled on a practice test, and no feedback is provided on the practice test, the distribution of memory strengths is "bifurcated," with the successfully recalled items receiving a large boost, whereas the nonrecalled items are unchanged (Kornell, Bjork, & Garcia, 2011). Thus, an immediate final test reveals essentially the same accuracy level as occurred on the practice test, whereas with restudy, there is an opportunity to learn items that were not initially acquired, resulting in higher accuracy. Furthermore, if restudy produces weaker learning, albeit learning that is applied to all items (i.e., a nonbifurcated distribution), and if memory strengths decrease at the same rate for all items with delay, then more memories of restudied items will fall below the retrieval threshold after a sufficient delay than will recalled items, reversing the pattern of results. This account suggests that accuracy on the practice test is a key factor, and when practice test recall accuracy is very high, there is an accuracy advantage for tested material over restudied material even for an immediate final test (Rowland & DeLosh, 2015).

Although a bifurcated distribution following test practice can explain this crossover interaction between delay and type of practice, it does not indicate *why* successful test practice produces more strengthening than restudy. Furthermore, this account is at odds with the finding that the crossover interaction holds even if the experiment uses only "retrievable" items, as indicated by the pretesting of all items (Jang et al., 2012).

One way to gain traction on the underlying mechanisms that support learning from recall practice is an examination of recall latencies. Latencies may be particularly informative in the case of a bifurcated distribution, allowing assessment of memory strength for the upper portion of the distribution (i.e., the items that were recalled on the practice test), considering that accuracy is nearly perfect for these items on the final test. Several studies have examined recall latency as a measurement tool to index retrieval effort during test practice (Pyc & Rawson, 2009; Vaughn, Dunlosky, & Rawson, 2016). A few others examined final recall latency following recall practice, finding that participants are faster to recall items after recall practice with free recall (Lehman, Smith, & Karpicke, 2014) and with cued recall (van den Broek, Segers, Takashima, & Verhoeven, 2014; also see Keresztes, Kaiser, Kovács, & Racsmány, 2014; van den Broek, Takashima, Segers, Fernández, & Verhoeven, 2013). In the case of the van den Broek study, participants were faster to recall items that had undergone recall practice as compared with restudy, even though accuracy was higher following restudy. This suggests that an analysis of latency may reveal attributes of the memory retrieval process not apparent when only considering accuracy. However, these studies only examined mean recall latencies, and we know from the application of sequential sampling models to latency distributions that a change in mean latency might reflect a change in response bias rather than an increase in the speed of the underlying process (Ratcliff, Thapar, Gomez, & McKoon, 2004; Ratcliff, Thapar, & McKoon, 2004). Thus, it is not clear whether these latency differences reflect a change in memory strength or whether they reflect a change in decision-making aspects of the recall paradigm (e.g., when to give up). The current study seeks to apply a decision-making model of reaction-time data to adjudicate between these theoretical alternatives.

## Primary and convergent retrieval

At this stage in its development, the primary and convergent retrieval (PCR) model, as first proposed by Hopper and Huber (2018), is a set of principles defining the learning that takes place during study or retrieval practice, and a sketch of the retrieval process necessary for recall. However, many computational details are currently unspecified, such as item similarity and the nature of the item features, the nature of stochastic noise during retrieval and learning, and the specific nature of context representations and changes in those representations with retention interval. Thus, no formal (mathematical) implementation of PCR exists at this time. Nevertheless, some qualitative predictions follow from the core principles of the PCR model regardless of these auxiliary assumptions, and the goal of the current study is to test these predictions to guide further development of the PCR model. This test of the PCR model requires application of a reaction-time decision-making model to rule out alternative explanations based on response bias.

In the PCR memory model, the primary retrieval and convergent retrieval processes are roughly similar to search and recovery processes in prior models (e.g., the sampling and recovery process in the SAM model: Raaijmakers & Shiffrin, 1981; the echo intensity and content responses in the MINERVA II model: Hintzman, 1984; and the cortical familiarity and hippocampal pattern completion processes in the complementary learning systems framework: Norman & O'Reilly, 2003). 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. In the primary retrieval stage of the PCR model, retrieval cues (e.g., explicitly provided item cues, or temporal context features) activate features of relevant target memories. However, this activation is incomplete for any particular item (i.e., some, but not all of the features are active). Successful recall requires convergent retrieval, a process whereby already active features excite inactive features based on associations between item features (i.e., intra-item associations). In brief, the PCR model contributes a greater consideration of retrieval pathways and the process of following along these pathways during the act of recall.

The PCR model assumes a feature representation, although the precise nature of these features is unspecific at this time. The features of a to-be-retrieved episode might, for instance, include perceptual, semantic, phonological, or orthographic attributes of the episode, and conjunctions of these attributes (see Criss & Shiffrin, 2004, for a similar proposal). This conjunction of attributes corresponds to the specific manner in which the item is interpreted during study (e.g., as seen in Fig. 1, upon encountering the word pair *table–bank*, one might create a mental image of a picnic on a river bank, and it is this conjunction of the lexical items and the created mental image that defines the episodic item). Although the features defining the episode are well known prior to the experiment, the particular conjunction is novel, and the pathways necessary for retrieval may not be well practiced (e.g., failing to remember that the correct answer is *bank* upon retrieval of the mental image of the picnic, particularly when considering that *bank* is more often a financial institution).

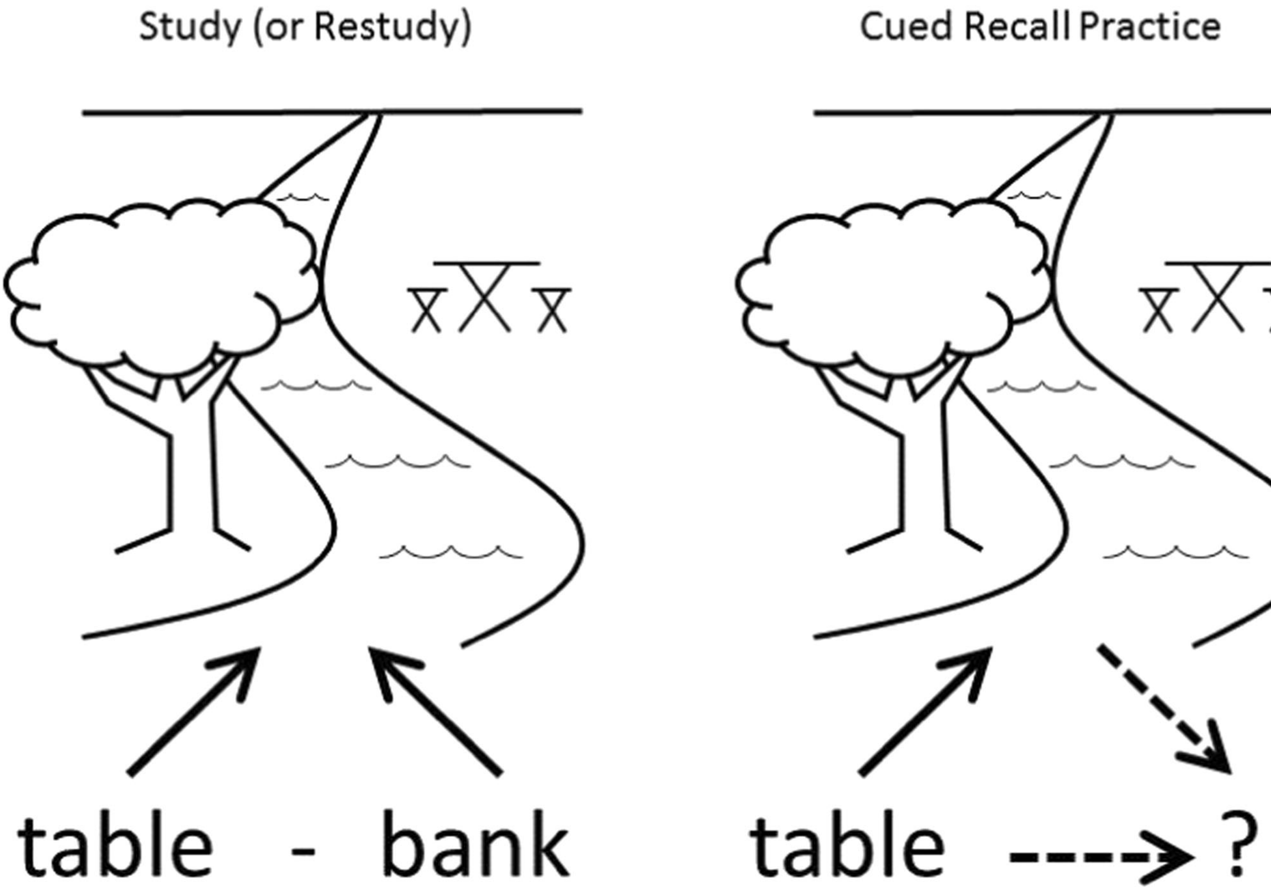

**Fig. 1** Different directed associations that are learned when studying (or restudying) the word pair *table–bank* (solid arrows) or when successfully recalling *bank* in response to the cue word *table* (dashed arrows). During initial study, a participant might create a mental image of a picnic table by a river bank to episodically conjoin the words. This produces directed associations from these words to the mental image. Restudy strengthens these forward associations, but this does not necessarily enhance recall of *bank* upon retrieval of this mental image, which relies backward associations. In contrast, successful cued recall practice involves activation of the mental image in response to *table* and then activation of *bank* in response to this mental image. According to the PCR memory model learning rule, this establishes directed associations from this mental image to the word *bank* as well as directed associations from *table* to *bank*, with both of these pathways boosting subsequent cued recall performance (both accuracy and latency) in response to the cue word *table*

The core assumption in the PCR model is that the associations between features are directional and are learned or strengthened depending on the temporal order of feature activation. More specifically, if two features become active at the same time, the association between them is not learned (or perhaps weakly learned), but if one feature becomes active and then the other becomes active shortly thereafter, the association between them is strengthened in a directed manner (from the first to the second). This assumption predicts that different types of practice will produce different kinds of learning, depending on the dynamics of the encoding event.

During restudy, the material is presented in its entirety (e.g., presentation of both *table* and *bank*). The temporal context is active before this presentation, allowing learning from that context to the presented material. In addition, the material may elicit memory for the original study episode (e.g., retrieval of the river picnic mental image), allowing additional learning from the presented material to the study episode. This boosts primary retrieval on a subsequent test (e.g., the context and any cue words on the final test more readily elicits the mental image). However, this boost may not prove to be helpful when the test involves recall rather than recognition. For instance, restudy may make it easier to recall the mental image of the riverside picnic in response to the cue word *table*, but this might only produce a modest boost for correct recall of *bank*, rather than some other aspect of the retrieved episode (e.g., *tree* or *river*).

In contrast to restudy, which activates features simultaneously owing to the representation of the material, recall success often occurs in a gradual staged fashion (Smith, Huber, & Vul, 2013). For instance, if asked to recall what you ate for dinner last night, you might arrive at an answer through a multistep process (e.g., “Yesterday was a Wednesday, and on Wednesdays my daughter has her ukulele lesson after school, so there’s little time to cook anything fancy . . . that’s right, it was pizza!”). This successful recall practice not only strengthens the associations supporting primary retrieval (e.g., from the context to the correct answer), it also strengthens associations between features within the episodically defined item (e.g., from the features of "Wednesday" to the features of "pizza"). We refer to this learning of associations between features of an episode as intra-item learning.[1] Intra-item learning protects an item from forgetting when activation from primary retrieval diminishes with an increased retention interval. For instance, contextual change between study and a later recall attempt (Mensink & Raaijmakers, 1988) makes it difficult to begin the process of recalling a dinner that occurred on January 2, 2019, after months or years have passed.

Critically, the temporal order of the stages during recall practice strengthens associations that support recalling the features of the episode in the same order on subsequent occasions. In the dinner recall example, the act of successful recall strengthens directed associations from "Wednesday" to "pizza," making it easier to recall this particular dinner menu at some later date by reconstructing the day of the week in question. Returning to the *table–bank* example, successful practice recalling *bank* in response to *table* strengthens the directed association from the mental image of the river picnic to the lexical entry *bank* (the backward-pointing dashed arrow in Fig. 1), bolstering recall based on this particular convergent retrieval path from *table* to *bank* via the mental image. In addition, because the correct answer is retrieved after the cue, this also supports a direct association from the cue to the answer, providing a more rapid retrieval path (the rightward-pointing dashed arrow Fig. 1).

This *table–bank* example is provided to aid intuitive understanding of the model, based on a particular assumption regarding the nature of the features that define an episodic item. Stepping back from this example, Fig. 2 provides a more abstract demonstration of the PCR model's predictions. These features might correspond to lexical entries and mental images, as in the *table–bank* example, but alternatively they might correspond to features *within* the lexical entry, such as when retrieving the first sound of someone's name (primary retrieval) and then attempting to recall the remainder of their name (convergent retrieval). The left-hand column shows a situation with successful recall practice. In this illustration, a feature becomes active if it has at least two lines of input from other active features (this particular activation rule is not a core assumption of the PCR model and is only adopted to spur intuitions about the model). Primary retrieval activates the first two features and then remaining features are gradually activated across four time steps. Because this is a gradual process, this produces intra-item learning, as indicated by the dashed arrows, which are directed associations from features that were active in earlier time steps to features that became active in later time steps. After this successful recall practice, a subsequent recall attempt starting from the same two initial features achieves full convergence in just two time steps owing to the new intra-item associations. Thus, the PCR model predicts that recall latencies will be faster after successful recall practice. Restudy might also produce faster recall by bolstering primary retrieval (i.e., a higher proportion of the features are active in the first time step). However, this speed-up owing to restudy should (1) be less than the speed-up after successful recall practice considering that recall practice boost primary retrieval and convergent retrieval, and (2) affect the latency distribution in a different manner than intra-item learning (we expand on this prediction below). In summary, because intra-item learning is an added benefit of recall practice, boosting previously unused retrieval pathways, the PCR model predicts faster recall following successful recall as compared with restudy, and this additional boost should affect the convergent retrieval process above and beyond a boost to the starting level of retrieved information.

Across both free recall and cued recall, Hopper and Huber (2018) confirmed the PCR model's prediction that recall practice reduces recall latencies more than restudy. Their free recall task measured accuracy and interretrieval time (IRT) for recall of 15-item word lists immediately following a practice recall test or a restudy. Restudy produced higher final test accuracy than did test practice, but accuracy and recall latency dissociated; an analysis of the IRTs revealed faster recall following recall practice, especially for the last few items that were recalled. The cued recall results reported by Hopper and Huber largely replicated the results of van den Broek et al. (2014), but additional conditions tested 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 with one of these cues. The final test cue was either the practiced cue or the unpracticed cue, and final recall was found to be faster only with the practiced cue, regardless of retention interval. This result is compatible with the directed nature of the learning rule in the PCR model; intra-item learning from a particular subset of features activated in response to a particular cue may not benefit convergent retrieval proceeding from a different subset of features activated in response to a different cue (i.e., learned navigation from one start point does not necessarily help navigation to the same goal from a different start point). This lack of generalization between retrieval cues argues against a context-based account of retrieval-based learning; if retrieval practice boosts associations between the temporal context and the target, then practice with one retrieval cue should transfer to another retrieval cue, provided that the temporal context is sufficiently similar between retrieval practice and the final test.

## 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

[1] The term "convergent retrieval learning" may also be used interchangeably.

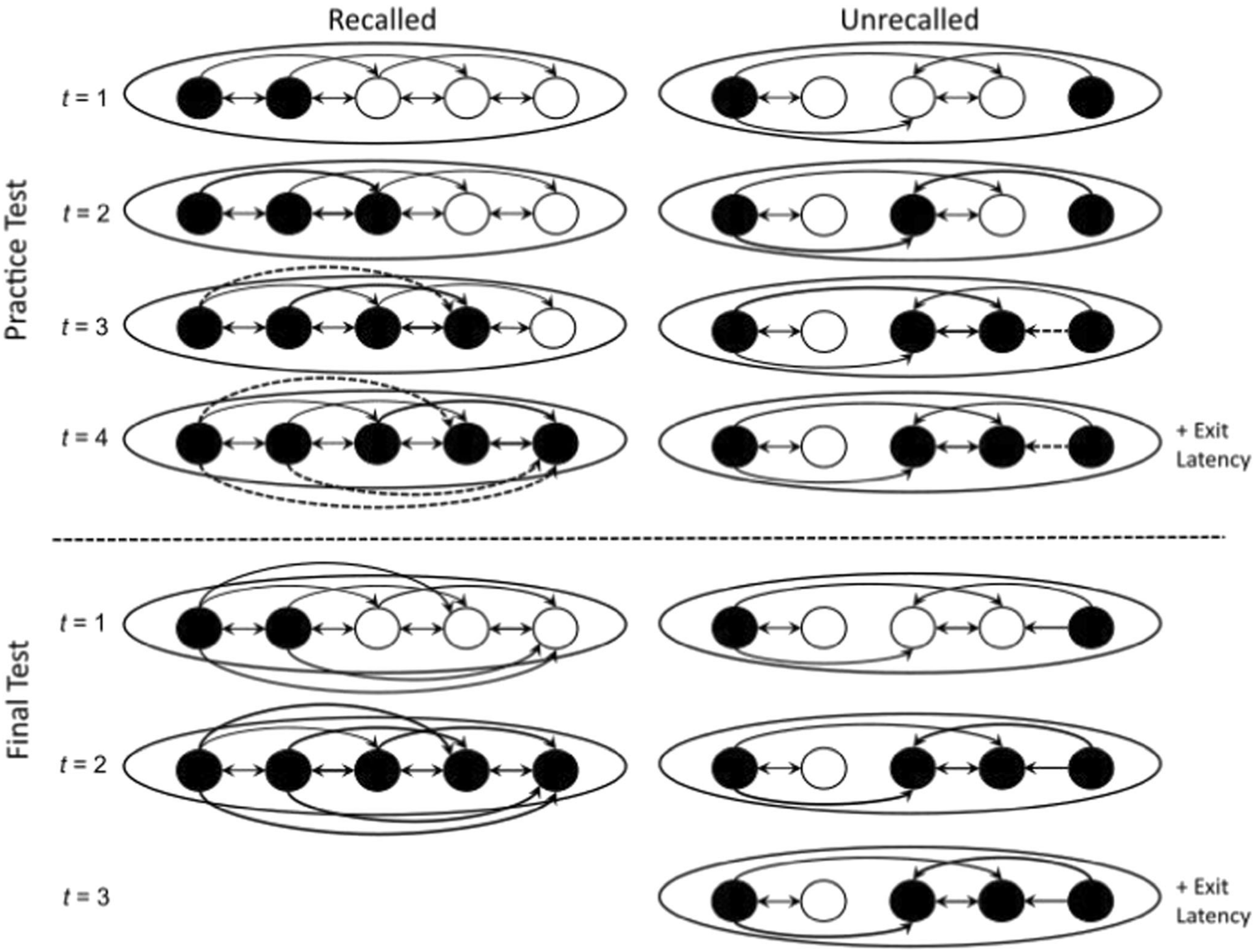

**Fig. 2** The convergent retrieval process and intra-item learning. Nodes represent episodic item features; filled nodes are active features, and unfilled nodes are inactive features. Arrowed paths represent associative connections. Solid arrows represent associative connections that activate subsequent features in the convergent retrieval process. The activation rule in this illustration requires two input connection from active nodes to cause an inactive node to become active. Dashed arrows represent new intra-item associative connections learned from retrieval. 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

predictions regarding the effects of recall failure, which are illustrated in the right column of Fig. 2. In this example, the item has a different set of preexisting intra-item connections, and these connections initially support the gradual activation of additional features. However, the convergent retrieval process ultimately reaches a dead end with no further change because 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).

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. We 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, we focus on recall failures (omissions) by requiring participants to make a binary "recall" or "can't recall" decision on each cued recall trial (see Fig. 3a). If the participant reports that they can recall the missing target, then 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 of 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 (2001) 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

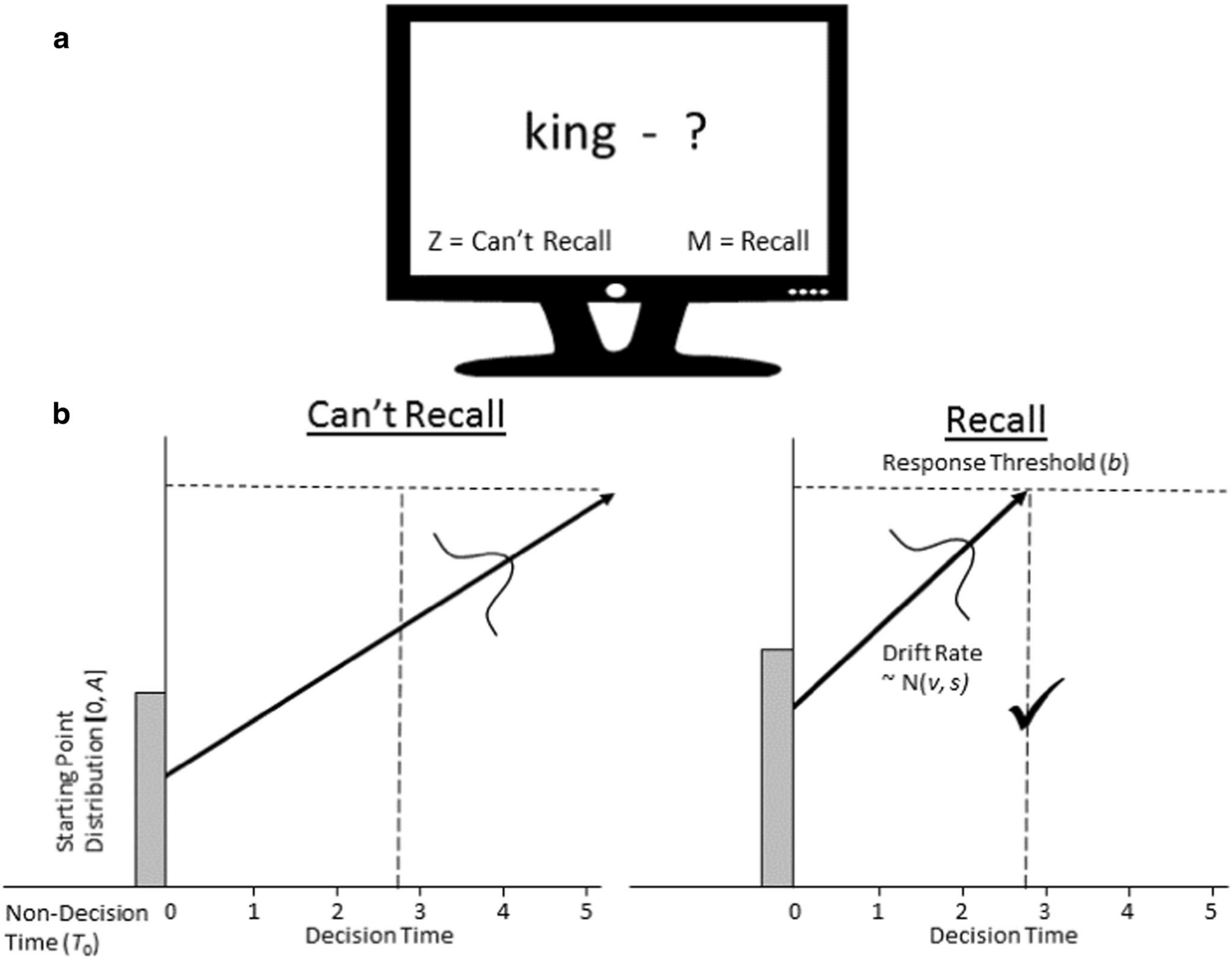

**Fig. 3** **a** Example of the "recall" versus "can't recall" decision made on each test trial of the current experiment. **b** Schematic representation of a decision between the "recall" and "can't recall" alternatives as described by the LBA model. The accumulator that intersects the response threshold 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

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.

## Recall latencies: A change in response bias (metamemory) 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 trade-off) 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" (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" versus "can't recall" decision. For example, Malmberg (2008a) showed that increasing the familiarity of a cue can increase the amount of time spent searching memory without increasing accuracy. Fortunately, we do not expect that cue familiarity will differ between the restudy and test practice conditions, given that both re-present the cue word.[2] 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. When available, fully implemented process models can be applied to decisional aspects of memory performance (Malmberg, 2008b; Malmberg & Xu, 2006; Malmberg, Zeelenberg, & Shiffrin, 2004), but this is not the case for the PCR model at this point in its development. 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 trade-off, and these models have proven useful in the study of memory (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. We follow in this tradition, using a sequential sampling model to transform our data into more psychological relevant parameters that, for instance, indicate whether the observed latency effects reflect the retrieval process (as predicted by PCR) or whether they reflect a change in the speed–accuracy trade-off (or more likely, some combination of these factors).

## 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, we 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 trade-off 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, she 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,

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

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 we 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, we could have applied a diffusion model 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 us to identify 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, we are not asserting that there are two independent convergent retrieval processes occurring simultaneously. Rather, we are using the accumulators of the LBA 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 Fig. 3b. 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 zero 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 nondecisional 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 preordained 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).

## 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—for example, is the stimulus a word or a nonword (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. Our 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, our goal with application of 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, we 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 Fig. 3b).

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 used 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.

We can distinguish between faster responding owing to a change in response bias versus 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 we use model comparison to determine which parameter changes provide the best account of the data.

To characterize the nature of any changes in retrieval latency and accuracy, we fit 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 of the LBA model should provide the best account of the data. We 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 with 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 of the 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 of the 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, we also considered hybrid models, allowing combinations of these parameters to vary across the experimental conditions.

## Overview of the current study

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 metamemory strategy to require less evidence to choose the "recall" response.

## Method

### 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.

### Materials

Twelve hundred (1,200) 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$_{\text{US}}$ 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

### 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 Fig. 4. This procedure yields a 2 × 3 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.

During study and restudy trials, word pairs were presented on the computer screen for 4 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 interstimulus 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.

## Results

### Statistical analysis

**Scoring** All latencies were measured as the duration between the onset of the cue word and the key press

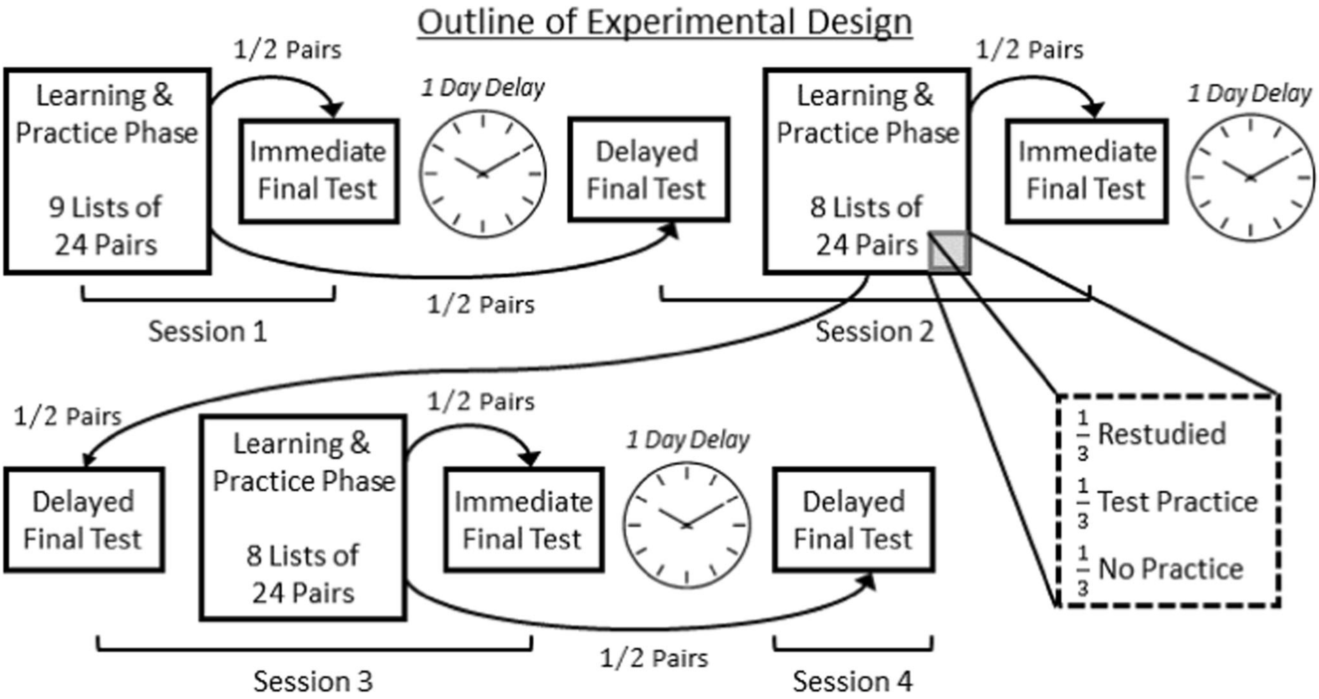

**Fig. 4** Outline of the experimental design, with the temporal structure of the sessions flowing left to right, and then top to bottom

indicating the recall/can't recall decision. 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.

**Recall accuracy** The percentage of words correctly recalled in each condition is shown in the top panel of Fig. 5. A trial was deemed correct only if the participant indicated they could recall the missing target word and subsequently typed the correct word. Differences in accuracy between conditions was 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.[3] 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 Fig. 5, it is clear that the main effect of retention interval reflects lower recall accuracy on the delayed final test. Differences in 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

[3] 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.

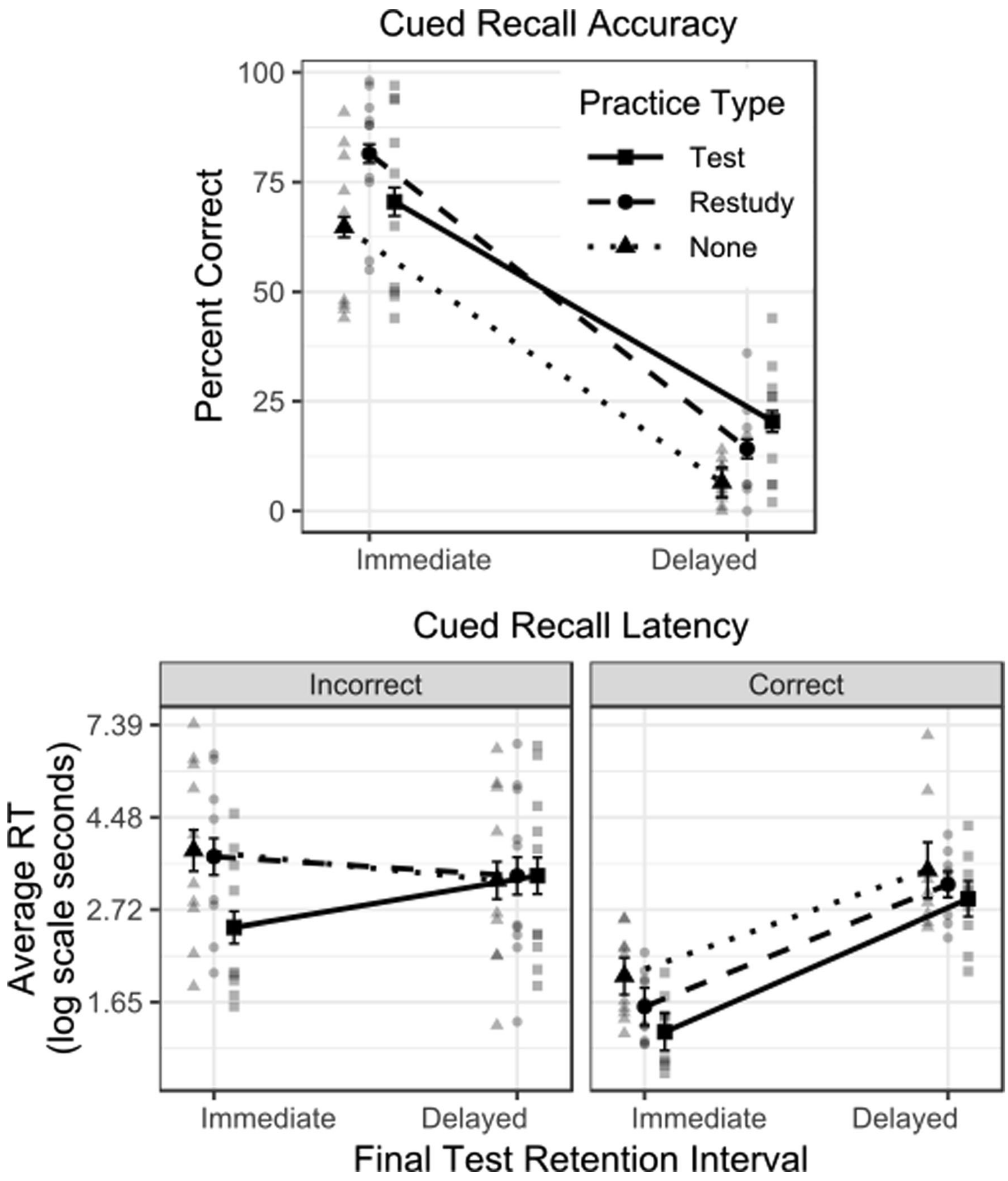

**Fig. 5** Performance across conditions. Larger, darker points represent averages across participants. Smaller gray points represent observations from individual participants. Error bars represent +/- one standard error of the mean, estimated using the subject-normalized method of Morey (2008). Top 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 indicated they could recall the target item, and subsequently provided the correct word as a response

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.

**Recall latency** The decision latency for “recall” and “can’t recall” judgements is shown in the bottom panel of Fig. 5. 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 Fig. 5, 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 “can’t recall” responses in the test practice condition on the immediate final test.

## LBA model analysis

**Overview** 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 model allowing the $b$ parameter to vary across conditions was the second best single parameter model (the $v$ parameter model was the best single parameter model), and so a sixth model was considered, allowing drift rate and boundary to vary ($v$ and $b$). The key question addressed by this sixth model was whether the predicted drift rate parameter value results would still hold when allowing boundary to change as well.

**Model details** 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[4] 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 (4 practice conditions [no practice, restudy, correct test practice, and incorrect test practice] × 2 retention intervals × 2 accumulators equals 16 drift rate parameters, plus two starting point parameters and the nondecision 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 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 that 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 allowed 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. This model allowed 33 free parameters per participant.

An initial analysis revealed that some participant's data were best described by a model with different boundaries across conditions, while data from other participants were best described by a model with different drift rates across conditions. To determine whether the drift rate parameter results would still hold when boundary was also allowed to vary, we fit a "*v* and *b* free" model that 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). This model allowed 33 free parameters per participant.

The actual number of free parameters 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.

**Model comparison** 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.

[4] 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).

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 ($\Sigma$ 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 ($\Sigma$ 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 data set 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.

**Model behavior** Beyond quantitative assessment of the models, we 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

**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 | 1652.62 | 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 |

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 our 10 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 with the no practice condition. To understand which free parameters were important in capturing the data, we performed multiple contrasts on the best-fitting parameter values using paired-samples $t$ tests.[5] 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, we 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 Appendix Tables 2 and 3. 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, Fig. 6 displays the joint quantile-probability plots for this model, showing the 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 $.6 \times .7 = .42$. The joint quantiles values were calculated for each participant individually and averaged together to create the values plotted in the figure.

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 with the "can't recall" responses, while the reverse is true in the observed data. However, *the model was not fit to these quantile distributions* and was instead fit to the raw trial-by-trial data (we only examine quantiles 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 crossover interaction, and matches the recall decision latency distributions reasonably well, we are satisfied that the "$v$ free" model describes the data accurately enough to interpret its parameter values.

[5] Identical $t$ tests were also performed for the "$v$ free" model, and described in detail in the Parameter Contrasts section.

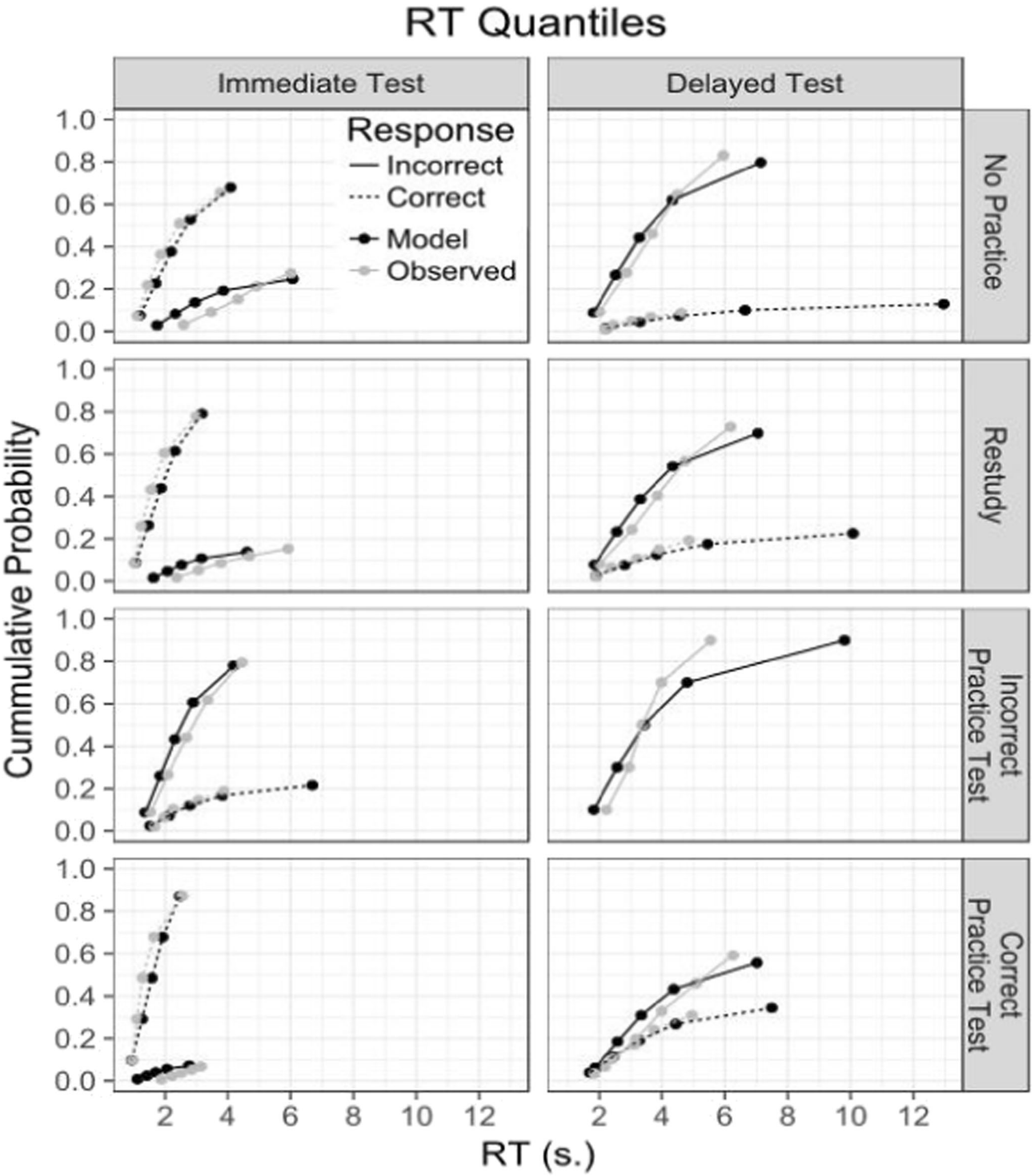

**Fig. 6** LBA Model RT quantiles, together with empirical quantiles estimated directly from the observed data. The quantile values were estimated at the .1, .3, .5, .7, and .9 quantiles of the RT distributions. No quantile functions for correct "recall" responses are presented for the delayed final test incorrect practice test condition because there were an insufficient number of responses for this situation

**Parameter contrasts** Because the model with different drift rate parameters in each condition was the most consistent winner, we 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 highest

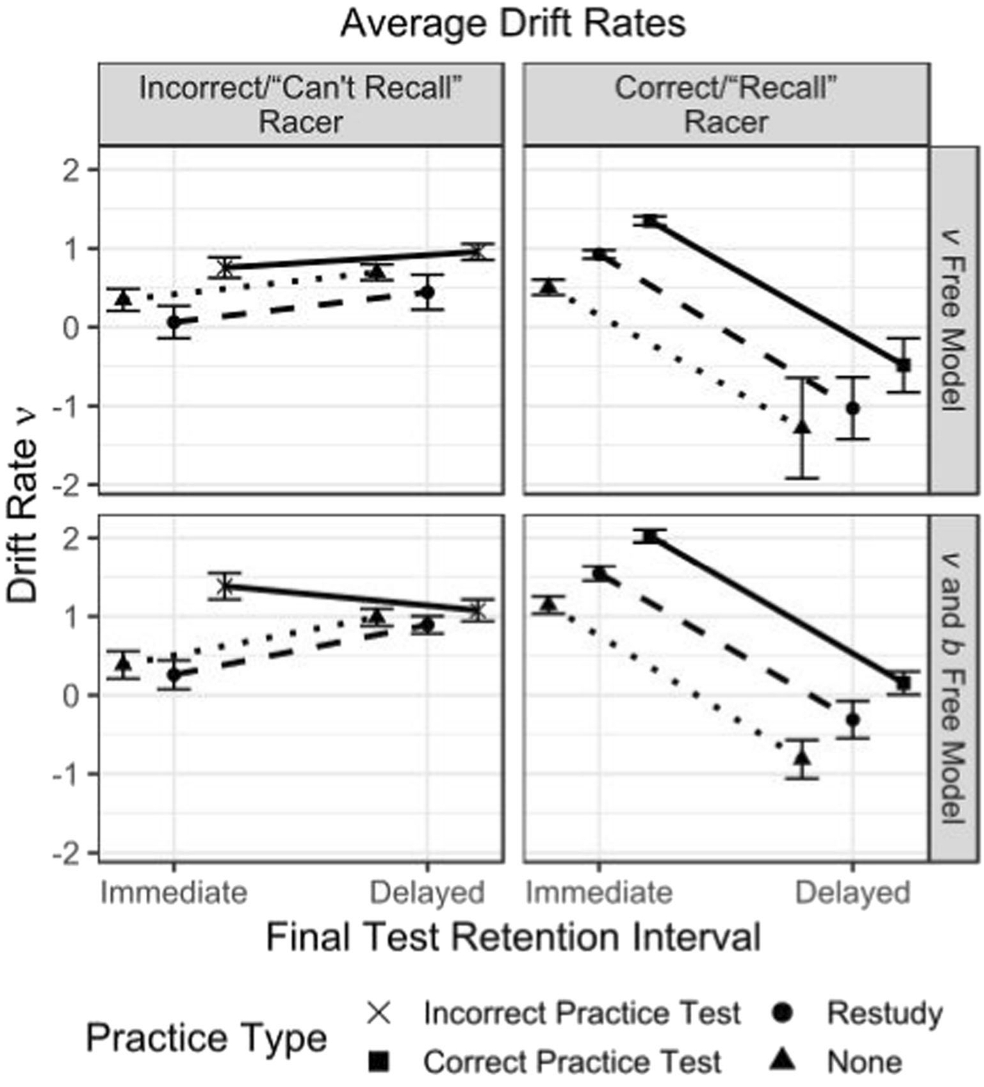

**Fig. 7** Average drift rates across participants for each condition from the "*v* free" (bottom row) and "*v* and *b* free" models (top row). Error bars represent +/- one standard error of the mean within each condition

after unsuccessful test practice. To test these predictions we 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 Fig. 7 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.[6]

First, we consider the "can't recall" accumulator's drift rate ($v_0$), shown in the lower-left panel of Fig. 7. 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 Fig. 6 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 Table 2 in the Appendix. 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 failure to recall.

[6] 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.

Next, we consider contrasts involving drift rates for the "recall" accumulator on the final test, shown in the lower-right panel of Fig. 7. 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 rate 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.

## Discussion

The Atkinson and Shiffrin (1968) modal model of memory, and subsequent implementations in the SAM and REM models, assume that recovery of memories is an all-or-none process, with the probability of retrieval proportional to the memory strength values dictating the search process. However, the tip-of-the-tongue phenomenon suggests that recovery is a gradual dynamic process that may ultimately result in success after an extended period of time, or might terminate after partial recovery with no further progress. The PCR model of Hopper and Huber (2018) assumes a dynamic recovery process, identifying learned associations that may be unique to recovery. Primary retrieval is the initial state of item retrieval based on the retrieval cues (temporal context and any presented cues). However, primary retrieval is often incomplete, and convergent retrieval is the recovery process in which already active item features activate inactive item features as dictated by directional intra-item associations between features. This account makes specific predictions regarding the time course of recall, and interactions between the time course of recall and the outcome of previous recall attempts. The current study confirmed these predictions by examining recall success and recall failure latency distributions as interpreted with a sequential sampling model of reaction times. We conclude that recall success speeds the rate of item recovery on a subsequent test, whereas recall failure speeds the failure to recover on a subsequent test.

The PCR model learning rule specifies directional associations 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 supports the learning of associations between the features of 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 with passive restudy. Previous work confirmed the prediction that these benefits result in faster recall even in situations where restudy produced better accuracy (e.g., for an immediate final test following restudy or following test practice without feedback). The current study confirmed that these recall latency benefits reflect the recall process itself (drift rate) rather than a metamemory response bias to hastily give a "recall" response based on knowledge that the prior recall attempt was successful. In addition, the PCR model predicted that intra-item associations are learned even if the convergent retrieval process stalls without reaching full convergence (i.e., learning from the failure to recall). The currently study confirmed these recall failure predictions 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; Nobel & Shiffrin, 2001). One reason for this is that a typical recall task does not ask participants to indicate *when* they have 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 (Nelson & Dunlosky, 1991), although in this case the judgment is immediately followed by typing in the answer if the "remember" option is chosen. The remember 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 (i.e., 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. By measuring both recall success and recall failure latencies, we were able to apply a reaction time decision making model (the LBA model) to these data to determine whether test practice caused a change in the speed–accuracy trade-off (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.

## Caveats and concerns

The PCR model makes no prediction as to whether subjects will or will not adopt a metamemory decision 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 as well as drift rate changes. As seen in the top two panels in Fig. 7, the pattern of drift rates is nearly identical for this model as compared with 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 indicated by the BIC measure, and as indicated 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. Our analyses separated the test practice condition into two pseudoconditions 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 pseudoconditions would select for easy versus hard items). We addressed this concern by directly comparing response latencies on the practice test to response latencies on the final test. We did this for two groups of items: Ones that were not recalled on either test and ones that were recalled on both tests. Providing clear evidence against an item selection effect account of these results, the average correct recall latency decreased by 1.22 seconds for a successful immediate final test as compared with 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 with recall failure latency for the same items on the practice test, $t(8) = 7.17$, $p < .001$.[7] 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 (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 (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, we applied the LBA model to the joint practice and final test data. As with the mean latency analysis, this was done for the 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 nondecision 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 pseudoconditions 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.

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

## Learning from failure

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, and 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 of 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 the first two letters 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 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.

## Sequential sampling model and recall decisions

Our study 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, we asked participants to judge the outcome of the retrieval process—whether it had identified the target word, or whether it had failed to recover the target. In our 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 that "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, our 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. (2014) 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. Our results are compatible with either explanation, and further experiments are necessary to understand the stopping rule used in cued recall tasks.

## Conclusions

Introspection suggests that recall is a gradual dynamic process that may stall (tip-of-the-tongue) or rapidly progress to the point that the desired memory can be named. This process was termed recovery by Atkinson and Shiffrin (1968), although the details of this process have remained largely unspecified during the 50 years since this seminal work. The PCR model of Hopper and Huber (2018) makes a specific proposal regarding the dynamics of recovery and the manner in which these dynamics enable additional learning from the act of recall. The current study tested key predictions of the PCR model that were confirmed with mean recall latencies and with a novel application of a sequential sampling model that ruled out alternative explanations based on response bias. These results move beyond the bifurcated distribution model of Kornell et al. (2011), specifying *why* a practice recall test promotes long-term learning. Furthermore, the two halves of the bifurcated distribution were separately evidenced, with recall success on a practice test resulting in faster recall on a subsequent test whereas recall failure on a practice test resulted in faster recall failure on a subsequent test.

## Appendix

**Table 2** Best fitting drift rates for the "*v* free" model

| Subject | Drift rate | Immediate test | | | | Delayed test | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | No practice | Restudy | Test practice, incorrect | Test practice, correct | No practice | 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 |

The 0 and 1 subscripts refer to the "can't recall" (incorrect) and "recall" (correct) accumulators, respectively. The "–" in some cells indicates that the parameter was not used for that participant

**Table 3** Best fitting starting point variability and nondecision time parameters for the "*v* free" model

| 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 |

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

## 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* (Tech. Report No. 79). Stanford, CA: Stanford University, Institute for Mathematical Studies in the Social Sciences. Retrieved from http://cogs.indiana.edu/FestschriftForRichShiffrin/pubs/1965%20Mathematical%20Models%20for%20Memory%20and%20Learning.%20Shiffrin,%20Atkinson.pdf

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation*, *2*, 89–195.

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

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and 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

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. https://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

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

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., & 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

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

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

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

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*(1), 1–67. https://doi.org/10.1037/0033-295X.91.1.1

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

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

Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: The role of retrievability. The Quarterly Journal of Experimental Psychology, 65(5), 962–975. https://doi.org/10.1080/17470218.2011.638079

Karpicke, J. D. (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*. New York: Elsevier. https://doi.org/10.1016/B978-0-12-809324-5.21055-9

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., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97. https://doi.org/10.1016/j.jml.2011.04.002

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., 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

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

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

Malmberg, K. J. (2008a). Investigating metacognitive control in a global memory framework. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory*. Abingdon, UK: Routledge. https://doi.org/10.4324/9780203805503.ch14

Malmberg, K. J. (2008b). Toward an understanding of individual differences in episodic memory: Modeling the dynamics of recognition memory. In A. S. Benjamin & B. H. Ross (Eds.), *Psychology of learning and motivation* (Vol. 48, pp. 313–349). New York, NY: Academic Press. https://doi.org/10.1016/S0079-7421(07)48008-2

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

Malmberg, K. J., & Xu, J. (2006). The influence of averaging and noisy decision strategies on the recognition memory ROC. *Psychonomic Bulletin & Review*, *13*(1), 99–105. https://doi.org/10.3758/BF03193819

Malmberg, K. J., Zeelenberg, R., & Shiffrin, R. M. (2004). Turning up the noise or turning down the volume? On the nature of the impairment of episodic recognition memory by midazolam. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 540–549. https://doi.org/10.1037/0278-7393.30.2.540

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

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial in Quantitative Methods for Psychology*, *4*(2), 61–64.

Nairne, J. S., Pandeirada, J. N. S., & Thompson, S. R. (2008). Adaptive memory: The comparative value of survival processing. *Psychological Science*, *19*(2), 176–180. https://doi.org/10.1111/j.1467-9280.2008.02064.x

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. https://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

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 [Computer software]. Vienna, Austria: R Foundation 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., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging*, *19*(2), 278–289. https://doi.org/10.1037/0882-7974.19.2.278

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.

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

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

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

Rowland, C. A., & DeLosh, E. L. (2015). Mnemonic benefits of retrieval practice at short retention intervals. *Memory*, *23*(3), 403–419. https://doi.org/10.1080/09658211.2014.889710

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., & 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 [Computer software]. Retrieved from https://CRAN.R-project.org/package=afex

Singmann, H., Brown, S., Gretton, M., & Heathcote, A. (2017). rtdists: Response time distributions [Computer software]. Retrieved from https://CRAN.R-project.org/package=rtdists

Smith, K. A., Huber, D. E., & Vul, E. (2013). Multiply-constrained semantic search in the Remote Associates Test. *Cognition*, *128*(1), 64–75. https://doi.org/10.1016/j.cognition.2013.03.001

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., Dunlosky, J., & Rawson, K. A. (2016). Effects of successive relearning on recall: Does relearning override the effects of initial learning criterion? *Memory & Cognition*, *44*(6), 897–909. https://doi.org/10.3758/s13421-016-0606-y

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

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