Cognitive Psychology 104 (2018) 106–142

Contents lists available at ScienceDirect

Cognitive Psychology

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

Modeling the dynamics of recognition memory testing with an integrated model of retrieval and decision making

Adam F. Osth[a,*], Anna Jansson[b], Simon Dennis[a], Andrew Heathcote[c]

[a] *University of Melbourne, Australia*
[b] *University of Newcastle, Australia*
[c] *University of Tasmania, Australia*

ARTICLE INFO

Keywords:
Recognition memory
Global matching models
Diffusion decision model
Testing effects

ABSTRACT

A robust finding in recognition memory is that performance declines monotonically across test trials. Despite the prevalence of this decline, there is a lack of consensus on the mechanism responsible. Three hypotheses have been put forward: (1) interference is caused by learning of test items (2) the test items cause a shift in the context representation used to cue memory and (3) participants change their speed-accuracy thresholds through the course of testing. We implemented all three possibilities in a combined model of recognition memory and decision making, which inherits the memory retrieval elements of the Osth and Dennis (2015) model and uses the diffusion decision model (DDM: Ratcliff, 1978) to generate choice and response times. We applied the model to four datasets that represent three challenges, the findings that: (1) the number of test items plays a larger role in determining performance than the number of studied items, (2) performance decreases less for strong items than weak items in pure lists but not in mixed lists, and (3) lexical decision trials interspersed between recognition test trials do not increase the rate at which performance declines. Analysis of the model's parameter estimates suggests that item interference plays a weak role in explaining the effects of recognition testing, while context drift plays a very large role. These results are consistent with prior work showing a weak role for item noise in recognition memory and that retrieval is a strong cause of context change in episodic memory.

A major constraint on models of memory concerns how the number of items present in memory affects memory performance. Such manipulations of memory set size have constrained models of recognition memory at both short (McElree & Dosher, 1989; Nosofsky, Little, Donkin, & Fific, 2011; Sternberg, 1966) and long (Clark & Gronlund, 1996; Dennis & Humphreys, 2001; Gillund & Shiffrin, 1984; McClelland & Chappell, 1998; Osth & Dennis, 2015; Shiffrin & Steyvers, 1997) time scales. Much theoretical interest concerns how the number of *studied* items in memory affects performance. However, of recent focus in recognition memory research is how the number of *tested* items affects memory.

Almost universally, recognition memory performance decreases throughout the course of testing. This finding was first reported by Peixotto (1947), but has frequently been replicated in the decades since (Annis, Malmberg, Criss, & Shiffrin, 2013; Averell, Prince, & Heathcote, 2016; Criss, Malmberg, & Shiffrin, 2011; Kiliç, Criss, Malmberg, & Shiffrin, 2017; Malmberg, Criss, Gangwani, & Shiffrin, 2012; Murdock & Anderson, 1975; Schulman, 1974). However, despite its status as an empirical regularity, theoretical interest in the nature of the testing effect has emerged only more recently (e.g.; Criss et al., 2011; Osth & Dennis, 2015). The decrease has been referred to as "output interference" in the literature, but we will reference it as the *test position effect*, or TPE, to avoid

[*] Corresponding author.
E-mail address: adamosth@gmail.com (A.F. Osth).

https://doi.org/10.1016/j.cogpsych.2018.04.002
Accepted 26 April 2018
0010-0285/ © 2018 Elsevier Inc. All rights reserved.

Table 1
Key terms and definitions used throughout the article.

Term	Definition
Test position effect (TPE)	Finding that recognition memory performance declines across test trials
List length effect	Finding that recognition memory performance is worse for longer study lists
Mirror effect	Effect of a manipulation that has opposite effects on hit rates (HRs) and false alarm rates (FAR). Low frequency (LF) words, for instance, have higher HRs and lower FARs than high frequency (HF) words (Glanzer & Adams, 1985)
Global matching model	Process model framework for recognition memory. Probe item is matched against all memories simultaneously; each similarity value is summed together to produce an index of memory strength that is used to make a decision
Context representation	Representation that defines a learning episode
Item noise	Interference generated by the studied items that are not the probe item (match in context, mismatch on item)
Context noise	Interference generated by occurrences of the probe item that were acquired prior to the study episode (match in item, mismatch in context)
Background noise	Interference generated by memories of items that are not on the study list (mismatch on item and context)
Context drift	Change in the context representation in response to events, such as study or test trials

commitment to the idea that the effect is driven by interference from items learned at test. In addition, we will demonstrate later in this article that changes in decision dynamics through the course of testing play a role in the observed phenomenon. Modeling of the TPE has explored causal factors acting through the decision process or memory retrieval, but not both. Our work aims to bridge this gap by introducing a combined model of memory retrieval and decision making that addresses the changes in both choice probabilities and response time (RT) distributions through the course of testing. Due to the large amount of nomenclature in the paper, a list of key terms can be found in Table 1.

1. Causes of the test position effect

One of the earliest attempts to explore the nature of the TPE was through Ratcliff (1978)'s application of the Diffusion Decision Model (DDM), an evidence accumulation model of the decision process (see Fig. 1). In the DDM, evidence begins at a starting point z and accumulates in a noisy fashion toward one of two response boundaries; an upper response boundary denoted by the parameter a (corresponding to a "YES" decision in recognition memory) and a lower boundary at zero (corresponding to a "NO" decision). The boundary first reached determines the choice made by the participant, while the time taken to reach the boundary, when added to the time for non-decision processes, determines the response time (RT). The phenomenon of the speed-accuracy tradeoff, whereby faster decisions are made less accurately, is captured by changes in the a parameter; increases in the boundary make errors less likely but increase the RT due to the longer distance that the process has to travel in order to reach a boundary. Memory strength in the model is conceptualized as the rate of evidence accumulation, or the *drift rate*; increases in the drift rate increase the proportion of correct responses and decrease the RT. Drift rate is not fixed but varies from trial-to-trial, which is analogous to cross-trial variability in memory strength in signal detection theory (SDT). Finally, non-decision time components, such as perceptual processing and response output, are modeled by parameter t_{ER}. Variability in non-decision time is assumed to have a uniform distribution with width s_t.

One might naively assume that changes in performance are due to changes in memory strength alone. However, according to the DDM, changes in performance across conditions can also be due to participants setting different speed-accuracy thresholds. Drift rate and speed-accuracy thresholds can be separably estimated in the DDM due to their differential effects on the RT distribution – increases in drift rate primarily decrease the skew of the RT distribution, while increases in response boundaries increase both the

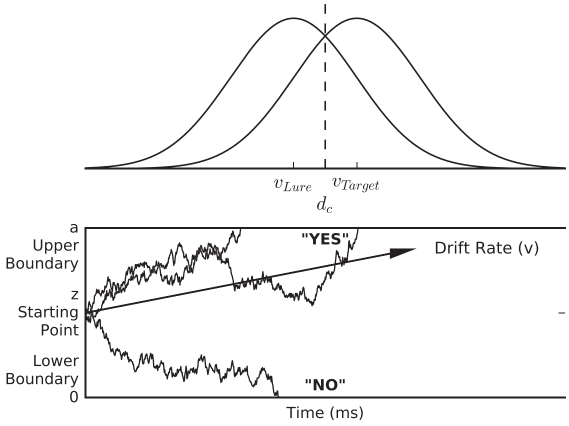

Fig. 1. The diffusion decision model (DDM). The drift rate is a sample from one of the normal distributions in the above panel. Evidence accumulation is noisy, such that diffusion processes with the same drift rate samples can reach different boundaries and produce different RTs. Depicted are three sample trajectories with the same drift rate. See the main text for more details.

leading edge (the earliest RTs) and the skew of the RT distribution (Ratcliff & McKoon, 2008). Ratcliff (1978) reasoned that drift rates, starting point, and response boundaries could all be affected during the course of testing, as participants could compensate for their decreases in accuracy by changing their threshold. Results of the modeling supported this conjecture; increases in test position were accompanied by decreases in drift rate, increases in response boundaries, and a decrease in the starting point toward the "no" boundary. These results suggest that recognition memory testing affects decision level components as well as memory strength.

Nonetheless, a weakness of the conventional DDM is that it is a *measurement model* of recognition memory, rather than a process model. Much like SDT, it is able to estimate distributions of memory strength but is agnostic as to the encoding and retrieval processes that generated them. The DDM is, therefore, unable to decompose drift rates into components that are meaningful within the context of theories of recognition memory, such as encoding strength, interference, or match to the episodic context. Without such a process model, one cannot know which sources are driving the decrease in drift rates over test trials in recognition memory.

Gillund and Shiffrin (1984)'s Search of Associative Memory (SAM) model, a global matching model of recognition memory, attributed the TPE to multiple causes. In SAM and other global matching models (see Clark & Gronlund, 1996, for a review), studied items are bound to a representation of the study episode or context representation. At test, the probe item and a representation of the study list context are jointly used to probe memory and matched against each of the stored item-context bindings, resulting in a single summed familiarity value that indexes the similarity between the cues and the contents of memory. It was considered a strength of the framework when it was first proposed that it naturally captures the list length effect, the finding that performance is worse after studying a longer list (Strong, 1912), because as the size of the list is increased, the variance of the familiarity distributions increases and the signal-to-noise ratio decreases.

Gillund and Shiffrin (1984) reasoned that the TPE has an analogous explanation; adding test items to memory through the course of testing increases interference. In essence, this is an assertion that the TPE is a list length manipulation that is occurring during the test phase instead of the study phase. We refer to this explanation as the *item noise account* of the TPE. The idea that test items are added to memory is an extremely reasonable, especially given that studies have demonstrated participants remember the lures tested in an experiment (e.g.: Jacoby, Shimizu, Daniels, & Rhodes, 2005). A question that remained was whether or not the interference from learning items at test is sufficient to explain the decline in performance. Gillund and Shiffrin also acknowledged that an additional possible cause of decreased performance was the increase in retention interval over the course of testing, which could, for example, cause a decreased match between the context cue employed at test and the stored context from the study episode. However, they could not distinguish between these two possibilities.

Criss et al. (2011) presented evidence against the retention interval account of the TPE. In their second experiment, participants were tested either immediately after the study list or after a 20 min delay. Performance was worse in the delayed condition, but contrary to the retention-interval account of the TPE, the decrease in performance through testing was much larger than the decrease in performance between the immediate and delayed conditions, despite the fact that the testing period took much less time than the delay period. The retrieving effectively from memory (REM) model (Shiffrin & Steyvers, 1997), which shares many features with the SAM model, provided a reasonable account of their data. However, they also acknowledged that models that lack item noise, such as the bind cue decide model of episodic memory (BCDMEM: Dennis & Humphreys, 2001), might be able to capture the results by assuming that the context representation changes as a direct consequence of recognition memory testing, an account which we will refer to as the *context drift* account of the TPE.

2. Evidence against the item noise account of the TPE

There are a number of grounds for doubting the item-noise account of the TPE. First, recent investigations of the list length effect have found that it is much smaller than previously believed and is even non-existent in several cases. Dennis and colleagues noted that there are a number of confounds present in list length designs that can artifactually induce a list length effect. For instance, if testing takes place immediately after the completion of the study list, the longer list has longer retention intervals on average than the shorter list. Second, attention likely decreases through the course of a study list due to boredom or fatigue (Underwood, 1978), reducing performance for later items on long lists. Dennis and colleagues recommended using controls such as filler activity upon completion of the study list that equates the retention intervals between the two conditions in addition to equating the serial position of the tested items by only testing beginning items. With these controls, list length was found to have either no effect (Dennis & Humphreys, 2001; Dennis, Lee, & Kinnell, 2008; Kinnell & Dennis, 2011; Kinnell & Dennis, 2012; Schulman, 1974) or a small effect (Cary & Reder, 2003; Kinnell & Dennis, 2012, for some non-linguistic stimuli) on recognition memory performance.

Schulman (1974) further demonstrated the lack of a list length effect when retention intervals were equated between study and test, but also found a very large TPE. The discrepancy between list length effects and TPEs is challenging for item noise models, which predict that both should co-occur. Nonetheless, there remains the possibility that memories stored during the test phase cause more interference than memories stored during the study phase, perhaps due to their greater recency. We will return to this potential issue later in the article.

Osth and Dennis (2015) argued against item noise accounts of forgetting on the basis of a model of recognition memory derived from the matrix model by Humphreys, Bain, and Pike (1989) that is able to measure the contributions of item and context noise. Osth and Dennis applied the model to ten recognition memory datasets from experiments using a wide range of manipulations including word frequency, study-test delay, list length, list strength, and stimulus class. The resulting parameter estimates suggested that item noise made at most a small contribution to forgetting in recognition memory.

Although Osth and Dennis (2015) estimated a very small amount of item noise, they did not model the TPE; all item noise was assumed to be generated by matching to memories acquired during the study phase. Nonetheless, they simulated the design used by

Schulman (1974) based on parameter values estimated from their fits along with an assumption that test items incremented item noise; this simulation indicated that the estimated magnitudes of item noise failed to generate enough interference to capture the TPE. They were, however, able to give a reasonable account of the Schulman data by implementing the context drift account. Context drift was introduced by Estes (1955) in the form of stimulus sampling theory, wherein context is represented as a set of elements that probabilistically change from one trial to the next. When context drift occurs during the course of testing, each test trial produces changes to the context cue used to probe memory, pushing it further away from the context of the study episode, and so hurting performance. With a small degree of context drift through the course of testing, the Osth and Dennis model was able to produce a substantial TPE while predicting a null list length effect, capturing the critical trends in the data.

Although trial-by-trial context drift has not generally been employed in recent models of recognition memory (but see Murdock, 1997, for an exception), it has a long history in models of episodic memory more generally (see Howard, 2014, for a review). Context drift was introduced into episodic memory models by Bower (1972), and was later used by Glenberg (1976) to unify the spacing and recency effects; the recency effect is predicted because the end-of-list context matches the final items on the study list, while the spacing effect is predicted because items that have been bound to many different contexts are more likely to match a context representation in the future than an item that is strongly bound to a single context.

It is usually assumed that context drifts not just during the events of the study episode, but during the events of the test phase as well. A number of empirical predictions have tested and confirmed the idea that the act of retrieval causes context drift to occur (Divis & Benjamin, 2014; Jang & Huber, 2008; Klein, Shiffrin, & Criss, 2007; Pastötter, Schicker, Niedernhuber, & Bäuml, 2011; Sahakyan & Hendricks, 2012; Sahakyan & Smith, 2014), a topic which we will return to in the General Discussion. This independent support for context drift in other areas of episodic memory research justify its inclusion in recognition memory models as a contender for explaining the TPE.

3. An integrated model of memory retrieval and decision making to explore causes of the test position effect

To summarize, Ratcliff's (1978) application of the DDM to the test position effect had the advantage of using both RT and accuracy data to decompose the changes in performance through testing into changes in memory strength, as measured by the drift rate, and changes in decision level phenomena such as bias and response caution, as measured by the starting point and response boundary. However, this approach was limited in that it was not able to decompose memory strength into factors that are relevant to theories of recognition memory, such as increasing interference from the items and changes in context through the course of testing. Existing recognition memory models describe such contributions but lack mechanisms to explain response time (but see Cox & Shiffrin, 2012), and thus variations in speed-accuracy thresholds across participants or across trials are erroneously attributed to memory processes.

We circumvent the limitations of each of these approaches by introducing a new model that uses the Osth and Dennis (2015) model to generate predictions about memory strength, which in turn are used to determine the drift rate distributions for the DDM, an approach which is quite similar to the exemplar based random walk (EBRW) model (Nosofsky et al., 2011). As discussed, the Osth and Dennis model has the ability to parameterize the degree of item noise and to parametrically implement assumptions regarding context drift. Thus, measurement of both of these parameters can be used to gain an understanding of which is more responsible for the performance decrements caused by recognition memory testing. We additionally measure changes in response bias and boundaries by parameterizing functional changes of each over test trials. This allows us to fit each test trial of the test sequence, providing higher resolution than previous work that fit blocks of contiguous recognition tests (Criss et al., 2011; Kiliç et al., 2017; Ratcliff, 1978).

The fundamental constituents of the Osth and Dennis (2015) model are items (I) and contexts (C) which are represented as vectors. The model stores conjunctive representations of items and contexts as outer products. These outer products are summed together to produce a memory matrix M that represents the sum total of learned experiences:

$$M = \sum_{t \in L} r(C_s \otimes \ I_t) \tag{1}$$

where r is a learning rate parameter. The subscript s indicates that the context vector corresponds to the study episode, subscript t indicates the item vector is an item from the list, and the set L corresponds to the items on the study list. Memory strength (s) is determined by combining the context cue along with the probe item cue at retrieval and matching it against the memory matrix M:

$$s = (C_s' \otimes \ I_t') \cdot M \tag{2}$$

where the dashes indicate that the cues employed may not be identical to the vectors stored at retrieval. Conventional applications of the matrix model proceeded by generating vectors from sampling distributions with a finite number of elements. The model circumvents this approach by using an approximate analytic solution that specifies the similarities between the vectors without specifying the content of the vectors. The derivations of the model and equations for the means and variances of the target and lure distributions can be found in Appendix A, with the concepts illustrated in Fig. 2. Similarities between the cues and the contents of memory are specified as normal distributions. The similarity between the item cue and the stored item vector is multiplied by the similarity between the context cue and the stored context vector. One can then easily derive means and variances of the target and lure distributions by summing the products of similarities across each item in memory.

Fig. 2 illustrates an example where the word "bubble" is used as a cue in conjunction with the study list context. Given that this was a studied item, there is a binding between "bubble" and a representation of the study list context present in memory. The match

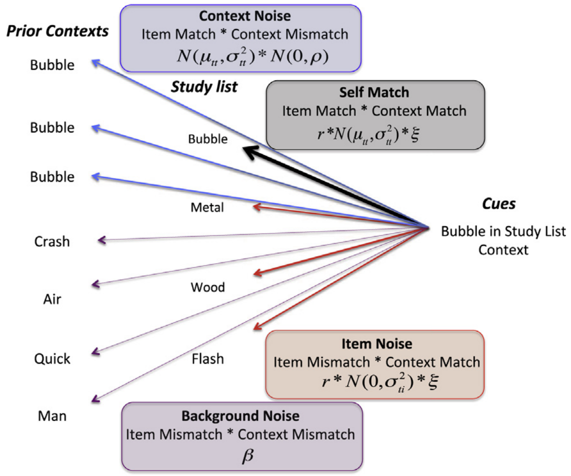
Prior Contexts
Context Noise
Item Match * Context Mismatch
$N(\mu_{tt},\sigma^2_{tt})*N(0,\rho)$
Bubble
Study list
Self Match
Item Match * Context Match
$r*N(\mu_{tt},\sigma^2_{tt})*\xi$
Bubble
Bubble
Bubble
Metal
Cues
Bubble in Study List Context
Crash
Air
Wood
Quick
Flash
Item Noise
Item Mismatch * Context Match
$r*N(0,\sigma^2_{ti})*\xi$
Man
Background Noise
Item Mismatch * Context Mismatch
β

Fig. 2. Diagram with an example of the Osth and Dennis (2015) model. Items in memory are associated with either the study list or prior contexts. A cue comprising an item cue ("bubble") and a context cue are globally matched against each item in memory. See the main text for details.

on the item dimension is a draw from a normal distribution with mean μ_{tt} and variance σ^2_{tt}, which are parameters of the model. The match on the context dimension has strength ξ. Both of these matches are multiplied together to produce an index of how similar the cues are to the stored conjunctive representation of "bubble" and the context learned at study.

Next, consider the match between the cues and the word "metal" in memory. There is a match on the context dimension, which has strength ξ, which is an additional model parameter that represents the effectiveness of the participant's contextual reinstatement. The stored item "metal", in contrast, mismatches the item cue "bubble." For this reason, the item similarity is a draw from the item mismatch distribution which has mean zero and variance σ^2_{ti}, an additional model parameter that controls the extent to which items interfere with other items on the list. The total similarity is a product of the context similarity and item similarity, which will have a mean of zero but nonzero variance. This calculation is done for all items stored in memory, and then their similarities are subsequently summed together to produce an index of how similar the cues are to the contents of memory. Because each item that is not the target item contributes variance, the signal-to-noise ratio is reduced as the number of items is increased or as the variance from each item match is increased.

The items in memory can be classified into four categories on the basis of whether there is a match or mismatch to the item and context cues (as illustrated in Fig. 2). A match on the item and context cue is called the *self match* and is the primary determinant of discrimination, as it is not present in the calculation of memory strength for lures. The mean of the self match is determined by the learning rate parameter r, the mean of the context match ξ, and the mean of the item match μ_{tt}. The learning rate r varies across conditions that differ in their study time or number of study presentations. The mean context match ξ varies across conditions that differs in their retention interval and also decreases with test position as a function of context drift (which is described later in the text). In our application of the model here the mean item match μ_{tt} was fixed to 1 in most conditions, but was allowed to take values less than 1 in a task-switching condition to represent deficient usage of the item cue.

The self match variance is primarily determined by the item match variability parameter σ^2_{tt}. Psychologically, this parameter corresponds to encoding variability of the probe cue, as the features present in the cue may vary from presentation to presentation (e.g., McClelland & Chappell, 1998). The self match term differentiates the variability of the target distribution from that of the lure distribution. If the item match variability σ^2_{tt} is greater than the item mismatch variability σ^2_{ti}, there is greater variability in the target distribution than the lure distribution, consistent with findings from recognition memory receiver operating characteristics (ROCs: Heathcote, 2003; Ratcliff, Sheu, & Gronlund, 1992; Wixted, 2007). RT data has reinforced these conclusions, as there is often greater variability in drift rates for targets than for lures (Osth, Bora, Dennis, & Heathcote, 2017; Starns & Ratcliff, 2014; Starns, Ratcliff, & McKoon, 2012). We will demonstrate in the General Discussion that the estimates of σ^2_{tt} were sufficient to produce estimates of target-to-lure variability that are comparable in magnitude to previous explorations with the DDM.

The *item noise* comprises variability in memory strength from items that are not the target item but were nonetheless present in the study list context. Increases in the number of items on the study list increase item noise. The level of item noise is primarily determined by the item mismatch variability parameter σ^2_{ti}; item noise increases monotonically with increases in σ^2_{ti} and if $\sigma^2_{ti} = 0$ there is no item noise. To model increases in item noise with recognition testing, each item adds a unique item noise term to Eqs. (9) and (10) (presented in Appendix A). Item noise variance increases linearly with increases in the number of items in memory. Since d' is measured in standard deviation units, item noise produces an inverse-square root relationship between the number of items in memory and d'.

The *context noise* comprises prior occurrences of the item cue in contexts that are outside of the study list context. Recognition memory experiments such as the ones in this article commonly employ words as stimuli, which have often times been experienced many times prior to the study list episode. These memories match the item cue but were learned in contexts prior to the experiment and mismatch the context cue employed at retrieval. Context noise is expected to increase monotonically with word frequency and is scaled by the parameter ρ, which varies across word frequency classes in our experiments.

The final term comprises the *background noise*, which is from items that were not present on the study list and were also not present in the study list context. Background noise comprises interference from all other unrelated memories acquired across the participant's lifetime.

3.1. The likelihood ratio transformation

Several of the model parameters, such as the context mismatch variability ρ and item mismatch variability σ^2_{ti}, only serve to increase or decrease the variances of the memory strength distributions and have no effect on the mean memory strength μ_{old}. The mirror effect is a finding where manipulations that increase performance produce opposite effects on the hit rate (HR) and false alarm rate (FAR; Glanzer & Adams, 1985), one example of which is that low frequency (LF) words exhibit higher HR and lower FAR than high frequency (HF) words. The mirror effect can be obtained if $\mu_{LFnew} < \mu_{HFnew} < \mu_{HFold} < \mu_{LFold}$.

This arrangement of distributions can be obtained via a log likelihood ratio transformation of the memory strengths, where the log likelihood ratio λ of a strength value x is $\lambda(x) = \log\left[\frac{f_{old}(x)}{f_{new}(x)}\right]$ (Glanzer, Hilford, & Maloney, 2009). We employ the linear approximation to the log-likelihood transformation devised by Osth, Dennis, and Heathcote (2017), which results in normal distributions. Expressions for means and standard deviations of the log likelihood ratio distributions can be found in Appendix A. Psychologically, the likelihood ratio transformation corresponds to the idea that memory strength alone does not drive a decision. Instead, each item's memory strength is compared to its expected strength before a decision is made. In a mixed list of strong and weak items, the expected strength for each item is an average of the weak and strong learning rates (e.g.; Starns, White, & Ratcliff, 2010).

After converting the memory strengths to log likelihood ratio distributions, a drift criterion d_c is applied to the log likelihood ratio distributions, which is an additional form of bias in the model. In contrast to the starting point z, which has a large effect on the leading edge of the RT distribution, the drift criterion shifts the drift rates of both targets and lures and thus primarily affects the skew of the RT distribution. Inclusion of the drift criterion d_c as a free parameter was partly motivated by the findings that there is substantial individual variation in bias across participants which has been found to reside in both the starting point and drift criterion (Bowen, Spaniol, Patel, & Voss, 2016). In many of our fits to data, the drift criterion was fixed across all conditions in the experiment.

3.2. Context drift

Implementation of context drift requires a function that relates context strength to the number of tested trials. We have used a modified variant of the Murdock (1997) contextual drift equation that easily allows for analytic derivation of the mean of the target distribution after context drift has occurred. Context drift introduces the parameter γ which reflects the proportion of context elements that are preserved from trial-to-trial. If $\gamma = 1$, no context drift occurs and the match to the study list context remains constant throughout the testing period. If $\gamma < 1$, context drift occurs and the mean of the context match decays exponentially throughout the course of testing.

Context drift occurs for both items that were stored at study and items that were stored at test. At test, we assume that items are bound to a context that is maximally similar to the current context (context match parameter $\xi = 1$ for tested items) and drifts in subsequent trials. This means that when $\gamma < 1$, recent test items are more similar to the current test context than older test items. The context match ξ scales the item noise (see Eqs. (9) and (10)) because the interfering items match the context cue, so any decrease in the match to the current context effectively decreases the item noise. This implies that items learned during recent test trials exhibit greater item noise than items learned during older test trials due to their context representations exhibiting a greater match to the context cue.[1]

In delayed testing conditions where testing begins with a partially reinstated context ($\xi_{delay} < 1$ for study items), the test trials exert more item noise than the trials stored at study by virtue of their stronger match to the context cue. This may explain why robust TPEs have been observed in cases where there was little to no effect of study list length (Schulman, 1974). Thus, while Osth and Dennis (2015) estimated very little role for item noise in their fits to data, the omission of items stored during the course of testing may have critically underestimated the degree of item noise. In all of our applications of the model, a single value of the context drift parameter, γ, was estimated for each participant, and was assumed not to vary across conditions.

Several models assume that context drifts through study trials as well as test trials (e.g.; Sederberg, Howard, & Kahana, 2008). In the present work we assume that context only drifts through the test trials for two reasons. First, our Experiment 1, which manipulates list length, tested target items in the same order in which they were presented, so that study-test lag was controlled. Second, the later datasets employed distracter activity before testing which, although brief (30 s), is often sufficient to eliminate effects of serial position in recognition memory tasks (Glenberg & Kraus, 1981; Talmi & Goshen-Gottstein, 2006). We do not mean to imply that context drift is not occurring through the course of study, but rather that the drift that is occurring may be negligible. In the General Discussion, we discuss additional evidence that retrieval may cause context to drift at a greater rate than during study trials.

[1] One should note that other assumptions are possible. For instance, participants might make a distinction between the contexts of study and test, with test items not having an association to the test contxt only. Under these circumstances, if participants's context cue was restricted to the study phase and not the test phase, test items would exhibit only minimal item noise despite the fact that they are being learned during the test phase. Although it is plausible, we have avoided using such an assumption as this would heavily bias the results of the modeling to favor a context-drift account of the TPE.

3.3. Changes in bias and speed-accuracy thresholds

Aside from the drift criterion, bias is additionally represented in the diffusion model with the starting point z. As with previous applications (Osth, Dennis, et al., 2017), we parameterize the starting point relative to the response boundary (z/a). We modeled changes in bias and speed-accuracy thresholds across the test trials as linear functions with slope and intercept parameters. Thus, unique values of z/a and a for trial i can be predicted with only four parameters: (z/a_{slope},z/a_{int},a_{slope},a_{int}). In contrast, Ratcliff (1978) divided the test sequence into blocks and allocated different values of z and a to each block. Our approach avoids the arbitrary division of the test sequence into blocks and uses fewer parameters. Although the choice of the linear function was itself only one of many possibilities, it can provide a reasonable approximation to many other smooth functions, at least over the limited test-position ranges we examined, and we found that it yielded sufficient fits to the data.

A common assumption in evidence accumulation models is that bias and threshold parameters can vary across conditions that are known to the participant, such as cross-list manipulations of retention interval, response proportions, or study list length (Donkin & Nosofsky, 2012b; Nosofsky et al., 2011; Ratcliff & McKoon, 2008). We follow suit here and vary both the slope and intercept parameters across some cross-list manipulations in our fits, such as list strength and the presence or absence of task switching, but hold the parameters fixed for within-list manipulations of stimulus difficulty.

Prior work with the DDM has used a uniform distribution of starting points with width s_z to capture the finding that errors are faster than correct responses under conditions of speed emphasis (Ratcliff & Rouder, 1998; Ratcliff, Van Zandt, & McKoon, 1999). We did not include start-point noise for several reasons because our prior work has found that starting point variability is unnecessary in fits to recognition memory data due to the absence of fast errors in recognition (Osth, Bora, et al., 2017).

4. Model predictions

To illustrate the effects of changing the chief parameters of interest – namely the item mismatch variability parameter σ_{ti}^2, which governs the total amount of item noise, the context drift parameter γ, and the speed-accuracy threshold change parameter a_{slope} – in isolation (i.e., while keeping the others constant) we generated predictions from the model where the parameter of interest had a non-zero value while the other two parameters were fixed to zero. Predictions are depicted in Fig. 3 for a list length paradigm identical to our Experiment 1, which consisted of a short list of 24 items and a long list of 96 items. To de-confound the effects of list length and test position, target items were tested in the same order as they were presented at study; this way, the first 24 test items of the long list had an identical number of test items in memory but only differed in the number of study items.

The predictions for each parameter are shown in a different column of Fig. 3. The predicted memory strength distributions (top row) and log likelihood ratio distributions (second row) are shown for the first trial (T1) of the short and long lists (blue and purple), along with the 96^{th} test trial (T96) of the long list (red) in order to illustrate the divergent effects of list length and test position. In particular, list length effects are revealed by a comparison of T1 for short and long lists, and test position effects are revealed by a comparison of the first and last test trial for long lists. For the context drift case (middle column), all lure distributions and the target distributions for the first test positions overlap due to the lack of item noise, with context drift only affecting the mean of the target distribution for the 96_{th} test position. All of the distributions overlap for the boundary manipulation (right column) because it has no effect on memory strength.

The log likelihood ratio distributions (second row) are used as inputs for the DDM, producing predictions for hit and false alarm rates (third row), correct response times (fourth row), and error response times (fifth row). The rate and RT predictions are shown as a function of contiguous sets of 24 trial (blocks), first for the entire short list (denoted “Short”), and then for 1_{st} to 4_{th} quarters of the long list (denoted 1, 2, 3, 4). Comparison of the first two values on these graphs (short vs. block 1) reveals the list length effect, while the change over the final four points (blocks 1–4) reveals the test-position effect. Response times are summarized using the 10^{th},50^{th}, and 90_{th} percentiles of the RT distribution; the 50^{th} percentile (median) measures the central tendency of RT, the 10^{th} percentile the leading edge of the RT distributions, and the 90_{th} percentile its long right tail.

For the predictions regarding item noise (left column) and context drift (middle column), all of the decision related parameters were kept constant across each condition, so effects are solely caused by differences in the log likelihood ratio distributions. Increases in item noise increase the variance of the memory strength distributions, while context drift decreases the mean of the target distribution. However, after the log likelihood ratio transformation, both of these parameters produce a mirror effect in the log likelihood ratio distributions, with the means of both the target and lure distribution moving closer together and having smaller variance as overall performance declines.

What both the item noise and context drift predictions have in common is large effects on the hit and false alarm rates along with increases in the skew of the RT distribution for correct responses with increasing test position. Error RTs, in contrast, are relatively constant across conditions. For instance, the item mismatch variability parameter σ_{ti}^2 predictions in the left column show a large effect of list length on the HR and FAR. Similarly, there are large increases in the .5 and .9 quantiles of the correct RT distributions with the increase in list length. One should note that for the item mismatch variability parameter σ_{ti}^2, the predicted test position effect is smaller than the list length effect due to the fact that adding more items to memory has a diminishing effect on performance due to the inverse square root function that relates the number of items in memory to d'. This means that the accrual of item noise has a non-linear effect on performance, with each item that is added into memory producing less and less of a detriment to memory performance.

The context drift parameter γ similarly predicts substantial effects of testing, as manifested in a decrease in HR and increase in FAR across long-list test blocks, along with a large increase in the .5 and .9 quantile of the correct RT distributions. What critically

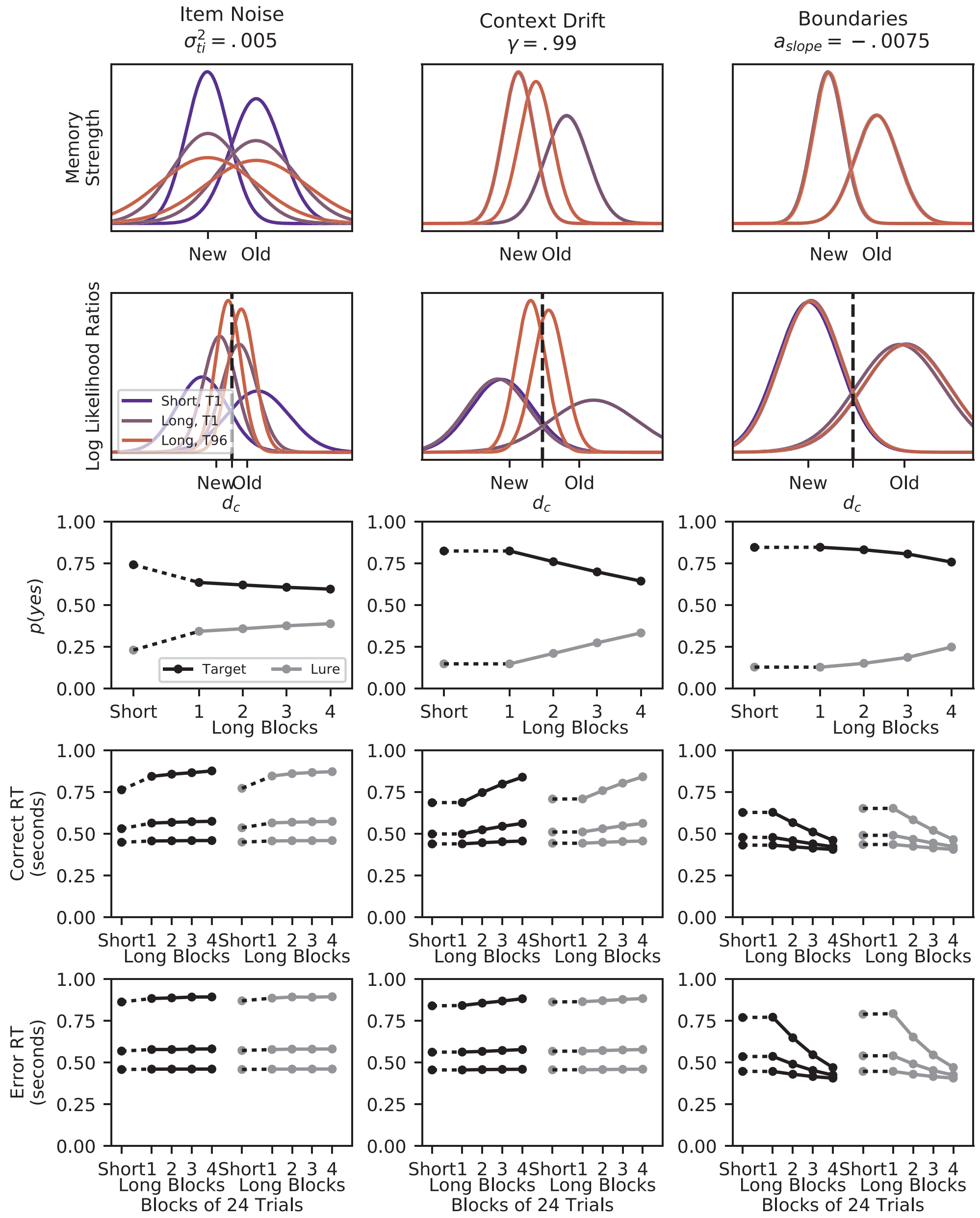

Fig. 3. Model predictions for each model parameter in isolation for a short list of 24 items and a long list of 96 items. The first column depicts the effect of increasing item noise through the item mismatch variability parameter σ^2_{ti}, the second shows the effect of context drift via the parameter γ, and the third shows the effect of decreasing response boundaries via the parameter a_{slope}. For each column, all other relevant parameters aside from the one of interest were set to zero. The top two rows depict the predicted memory strength and log likelihood ratio distributions for the first trial of the short and long lists along with the 96th trial of the long list. HR and FAR (third row), correct RTs (fourth row), and error RTs (fifth row) are depicted for the short list ("Short" on the x-axis) and quarters of the long list (1, 2, 3, 4 on the x-axis). All other model parameters: $\sigma^2_{tt} = .1, r = 1.0, \xi = 1.0, \beta = .05, d_c = 0, a_{intercept} = 1.0, z/a_{intercept} = .5, z/a_{slope} = 0, t_{ER} = .4, s_t = 0$. To equate performance on the short list across the conditions, ρ was equal to .02 for the item noise plot and .05 for the context drift and decreasing response boundary plots.

distinguishes the context drift parameter γ from the item noise predictions is that no effect of list length is predicted; performance for the short list and the first block of the test list are predicted to be identical if decision-related parameters (such as bias and speed-accuracy thresholds) are held constant across those conditions. One should note that the model predicts no list length effect only because study-test lag is equated across the two conditions. If testing were to take place immediately after the study list, context drift can predict list length effects because study list items in a short list are associated with contexts that are more similar to the context cue than the contexts associated with early items in the long list.

Decreases in speed-accuracy thresholds contrast with the other two sets of predictions in two important ways. First, the impairments in HR and FAR through the course of testing are relatively small compared to the changes in RTs. Second, RTs *decrease* as testing continues, rather than increase. Finally, error RTs are affected to a larger degree than correct RTs. The final prediction is perhaps the largest distinction from the other two parameters, which predict very little change in the error RTs. Note that while Fig. 3 depicts the case where speed-accuracy thresholds decrease through the course of testing, the opposite trend is also possible, which likewise makes the opposite predictions: increases in accuracy and RTs through the course of testing.

Fig. 3 depicts simplified predictions from each of the components in isolation, but in practice each component may be present and will interact in the full model. A relevant consideration is that Fig. 3 shows a mirror effect during recognition testing (decreasing HR and increasing FAR), while in practice the FAR are often unaffected by testing (Criss et al., 2011) or may even decrease (Koop, Criss, & Malmberg, 2015). It is worth noting that when there is bias present in the model (either via starting point or drift criterion) or changes in bias through testing, the model makes different predictions about the FAR pattern. For instance, if bias becomes progressively more conservative through the course of testing, this will decrease the FAR and mitigate the FAR predictions of the other parameters. We will later demonstrate that the model is capable of accommodating an approximately flat FAR pattern in our fits to data.

In addition, because the TPE is likely the product of a number of memory and decision related factors, it can be difficult to understand which parameters are primarily responsible for the effect. Throughout the article, we clarify the role of each parameter by generating TPE predictions and correlating them with each of the parameters.

5. Hierarchical Bayesian modeling

The model was applied to data using hierarchical Bayesian methods (see Lee, 2011; Rouder & Lu, 2005, for introductions), which offer a number of advantages over conventional model fitting methods. First, knowledge about the model parameters can be used to constrain the parameter space through specification of a prior distribution, which can constrain the model to behave in a manner consistent with psychological theory (e.g.; Vanpaemel & Lee, 2012) or to inherit constraints from prior applications to data (e.g.; Kary, Taylor, & Donkin, 2016). In subsequent applications of the model, we use informed priors based on the posterior estimates from the first experiment to accomplish both purposes. Specifically, Experiment 1 contains a list length manipulation, which disentangles the predictions of item noise and context drift for the TPE. Using priors informed by the results of the first experiment allows us to inherit that constraint for the later datasets that lack such a manipulation. In addition, there is no a priori theoretical reason why the parameters governing memory retrieval should change across datasets, and thus informed priors allow us to preserve such a consistency. Decision-related parameters, in contrast, can change depending on the nature of the test instructions and perceived difficulty of the experimental task, and thus were given relatively uninformative priors for each application. Our conclusions did not change when uninformative priors were employed. We present these results in the Prior Sensitivity Analysis in the Supplementary Materials.

In addition, hierarchical methods are advantageous in estimating data from individual participants when there are not a large number of trials per participant; this is because in hierarchical models estimates of the participant level parameters are influenced by the group level parameters. When there is a large degree of uncertainty in the individual parameter estimates, they get pulled toward the group estimate, a phenomenon referred to as “shrinkage”. This is advantageous in fitting the present model as the DDM requires large numbers of trials per participant to reliably estimate its parameters (Wagenmakers, 2009). The advantages of hierarchical modeling of the DDM were also detailed by Vandekerckhove, Tuerlinckx, and Lee (2011).

Estimation of the posterior distribution requires Markov chain Monte Carlo (MCMC) algorithms. However, in process models parameter estimates are often correlated with each other (Ratcliff & Tuerlinckx, 2002; Turner, Sederberg, Brown, & Steyvers, 2013), which is problematic for conventional MCMC algorithms. For this reason, we used differential evolution Markov chain Monte Carlo (DE-MCMC: Turner, Sederberg, et al., 2013), a method of posterior sampling that is robust to parameter correlations. We encourage interested readers to consult the Turner, Sederberg, et al. (2013) article for a detailed and technical description of this procedure.

For Experiment 1, relatively non-informative priors were employed on the model parameters; these prior distributions and details of the fitting procedure can be found in Appendix B.

Due to the large number of possible model parameters and identifiability issues when applying the model to a single dataset, we applied sensible restrictions by fixing parameters whenever possible. A complete list of all model parameters for the present experiment can be found in Table 3. The Supplementary Materials contain the results of a parameter recovery where we demonstrated that we were able to recover the parameters of the model for most of the participants in each dataset.

6. Overview of experiments

We have described a process model of recognition memory that is able to make predictions about both choice and response times by using a global memory matching front-end with a back-end diffusion process for the decision stage. Through the course of this article, we apply the model to four datasets in an attempt to explore the magnitudes of each of the potential causes of the TPE, including increasing item noise, context drift through the coures of testing, and changes in speed-accuracy thresholds. The first is an experiment we conducted that manipulated list length to evaluate whether the model is capable of producing a TPE with relatively little effect of list length. This dataset is ideal in dissociating the effects of context drift and increasing item noise, as item noise is jointly constrained by the effects of list length and test position while context drift does not predict effects of list length.

To forecast our findings, the results of Experiment 1 indicated a very strong role of context drift in driving the TPE. In order to further constrain this account, we subsequently applied the model to three datasets that have been previously argued to challenge a context drift account. The second and third datasets employ cross-list and between-list manipulations of strength, respectively. We

Table 2
Summary of the datasets fit by the model.

Dataset	N	Obs.	LL	TL	Conditions
Experiment 1	107	233	24/96	24/96	List length (24/96, cross-list), RI (immediate/delayed, cross-list), WF (LF/HF)
Criss (2010, E2)	16	1519	50	100	Presentations (1 × /5 × , cross-list), WF (LF/HF, cross-list)
Starns (2014, E2)	33	290	50	100	Presentations (1 × /3 ×)
Annis et al. (2016, E2)	59	309	80	160	Between-trial activity (blanks/LD, cross-list)

Notes: E = experiment, N = number of participants, Obs. = mean number of observations per participant, LL = study list length, TL = number of recognition test trials, Conditions = independent variables manipulated, RI = retention interval, WF = word frequency, LF = low frequency, HF = high frequency, LD = lexical decision.

chose these manipulations based on recent evidence showing that the TPE is smallest in lists composed entirely of strong items, equal for weak and strong items in mixed lists, and steepest in lists composed entirely of weak items (Kiliç et al., 2017). The final dataset employs a task-switching manipulation, which we chose based on evidence that interspersing lexical decision trials between test trials in recognition memory does not worsen performance despite a functional doubling in the number of test trials that participants experience (Annis et al., 2013).

Each of the datasets employ relatively long test lists, which provides constraint on the parameters related to the dynamics of testing. A table documenting the number of participants, data points per participant, length of each test list, and conditions in the experiments can be found in Table 2. Each of the datasets employed yes/no testing. While this is contrary to Criss et al. (2011)'s recommendations of using two alternative forced choice (2AFC) tests, recent evidence measuring RTs (Jou, Flores, Cortes, & Leka, 2016) and eye movements (Starns, Chen, & Staub, 2017) has cast doubt on whether participants in 2AFC testing use relative judgments, making the DDM an appropriate choice, or absolute judgments, which would make a race or accumulator model more appropriate (e.g.; Brown & Heathcote, 2008; Usher & McClelland, 2001).

7. Experiment 1: list length, word frequency, and study-test delay

Our experimental design manipulates list length (between 24 and 96 items), word frequency (LF and HF words), and study-test delay, to constrain the relevant parameters of our model. The design also uses several of the controls advocated by Dennis and colleagues for confounds present in list length designs. Specifically, retention intervals between the short and long list were equated by (a) testing items in the same order in which they were studied and by (b) using filler activity to equate the time at which the test list begins between the short and long lists. Unlike previous designs which only tested the beginning items of the long list, all serial positions of the long lists were tested. In order to minimize the contributions of rehearsal, participants made pleasantness ratings to each item during the study list.

7.1. Method

7.1.1. Participants

One-hundred and seven participants from the University of Newcastle participated in exchange for course credit.

7.2. Materials

Words were drawn from the Google word frequency counts, which provides a measure of how frequently words are used on various websites on the internet. Four-hundred and eighty words between 3 and 11 letters in length and either between 1 and 4 counts per million (low frequency, or LF) or 100–200 counts per million (high frequency, or HF), were used in this experiment.

7.3. Procedure

The study phase comprised either short lists of 24 items or long lists of 96 items with an equal number of LF and HF words in each study list. During the study phase, each item was displayed for two seconds individually in the center of the screen in white uppercase font on a black background. Participants were asked to rate their pleasantness on a 4 point scale from "very unpleasant" to "very pleasant" and make their responses on a keyboard on keys 1 through 4, respectively. To ensure that participants could remember the keys for each response option, the response keys remained on the screen through the duration of the trial.

The study lists in the delayed condition, along with the short list in the immediate condition, were followed with filler activity, a digital card game. Playing cards were presented on the center of the screen one at a time and participants were instructed to press keys on the keyboard when various rules were met, such as pressing the spacebar when two cards in a row had the same suit, or pressing the "j" key when they saw the joker. To motivate participants to engage with the filler task, a running tally of the score was presented with points being given for correct responses and deducted for incorrect responses.

Test lists were of the same length as the study lists. To provide more control over study-test lag, targets were presented in the same serial position as in the original study list. Half of the study list items were randomly selected to be presented on the test list while the remaining empty positions were substituted with lures. For example, if a study list ABCDEF was studied and items A, B, and F, were

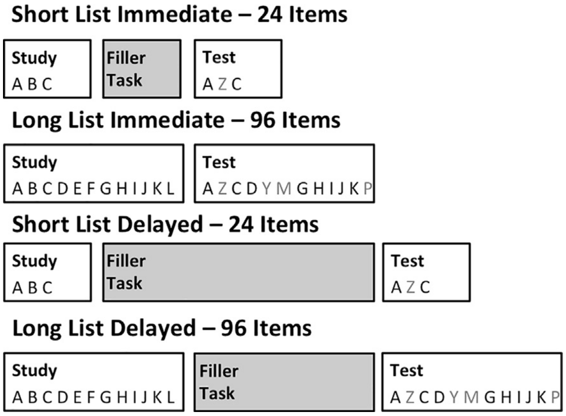

Fig. 4. Diagram of the experimental procedure. Example study lists and test lists with reduced numbers of items are depicted, with studied items in black and lure items in gray. Note that in the test lists the targets are presented in the same order as on the study list.

randomly selected to be tested as targets, an example test list would be ABXYZF. During the test phase, participants were presented with words one at a time on the center of the screen and were asked to press "1" if they recognize the word from the study list and "0" if they did not recognize it. The response options remained on the screen for the duration of each trial to ensure that participants could remember the response keys. A diagram depicting the experimental conditions along with example study and test sequences can be found in Fig. 4.

7.4. Model parameterization

We fixed the learning rate r to 1 for simplicity in the present application (although later fits allow the parameter to vary across conditions varying in the number of presentations). In the immediate testing conditions, ξ for study items was fixed to 1 for the first trial to reflect a strong match to context. In delayed conditions, the value of ξ for study items on trial 1 was estimated as a free parameter ξ_{delay} to reflect the weaker contextual match after the larger retention interval. In the Supplementary Materials, we present the results of a Parameter Robustness Analysis where we applied models with different values for these parameters, and each of them demonstrated similar results.

Because a list length manipulation is a cross-list manipulation, we allow the bias and threshold intercepts $z/a_{int.}$ and $a_{int.}$ to vary across the short and long lists (cf. Ratcliff & McKoon, 2008). In addition, increasing retention intervals have been found to produce an

Table 3
Model parameterizations and number of parameters per participant (N) for each dataset.

Param	Description	Bound	Exp. 1	Criss	Starns	Annis
r	Learning rate	0: ∞	$r = 1$	R	R	1
ξ	Context match	0: 1	$\xi_{imm} = 1, \xi_{del}$	$\xi = 1$	$\xi = 1$	$\xi = 1$
μ_{tt}	Item match	0: 1	$\mu_{tt} = 1$	$\mu_{tt} = 1$	$\mu_{tt} = 1$	$\mu_{tt,bl} = 1, \mu_{tt,LD}$
σ_{tt}^2	Item match variability	0: ∞	1	**1**	**1**	**1**
ρ_{avg}	Context mismatch variability	0: ∞	1	$\rho_{avg} = .039$	$\rho_{avg} = .039$	$\rho_{avg} = .039$
ρ_{wf}	Proportional diff. in ρ between HF/LF	0: 1	1	**1**	–	–
σ_{ti}^2	Item mismatch variability	0: ∞	1	**1**	**1**	**1**
γ	Rate of context drift	0: 1	1	**1**	**1**	**1**
β	Background noise	0: ∞	$\beta = .05$	$\beta = .05$	$\beta = .05$	$\beta = .05$
d_c	Drift criterion	−∞: ∞	1	1	1	T
z/a_{int}	Starting point	0: 1	L,D	R	1	T
z/a_{slope}	Change in z/a through testing	−∞: ∞	1	R	1	T
a_{int}	Response boundary	0: 1	L	R	1	T
a_{slope}	Change in a through testing	−∞: ∞	1	R	1	T
t_{ER}	Mean nondecision time	0: ∞	1	1	1	T
s_t	Nondecision time variability	0: ∞	1	1	1	T
N	–	–	17	17	12	19

Notes: Letters indicate the experimental factor the parameter was varied over, "1" indicates a single parameter was estimated across all conditions, bold entries indicate parameters that were estimated using informed priors from the previous fit (the column to the left), entries that show the parameter name set equal to a value denote fixed parameters, while entries that show a single parameter name in a condition were estimated for that condition only. Entries with a "–" indicate the parameter did not apply to that dataset. E = experiment, L = list length, D = delay, F = word frequency, R = repetitions, LD = lexical decision trials condition, imm = immediate condition, del = delayed condition, bl = blank condition.

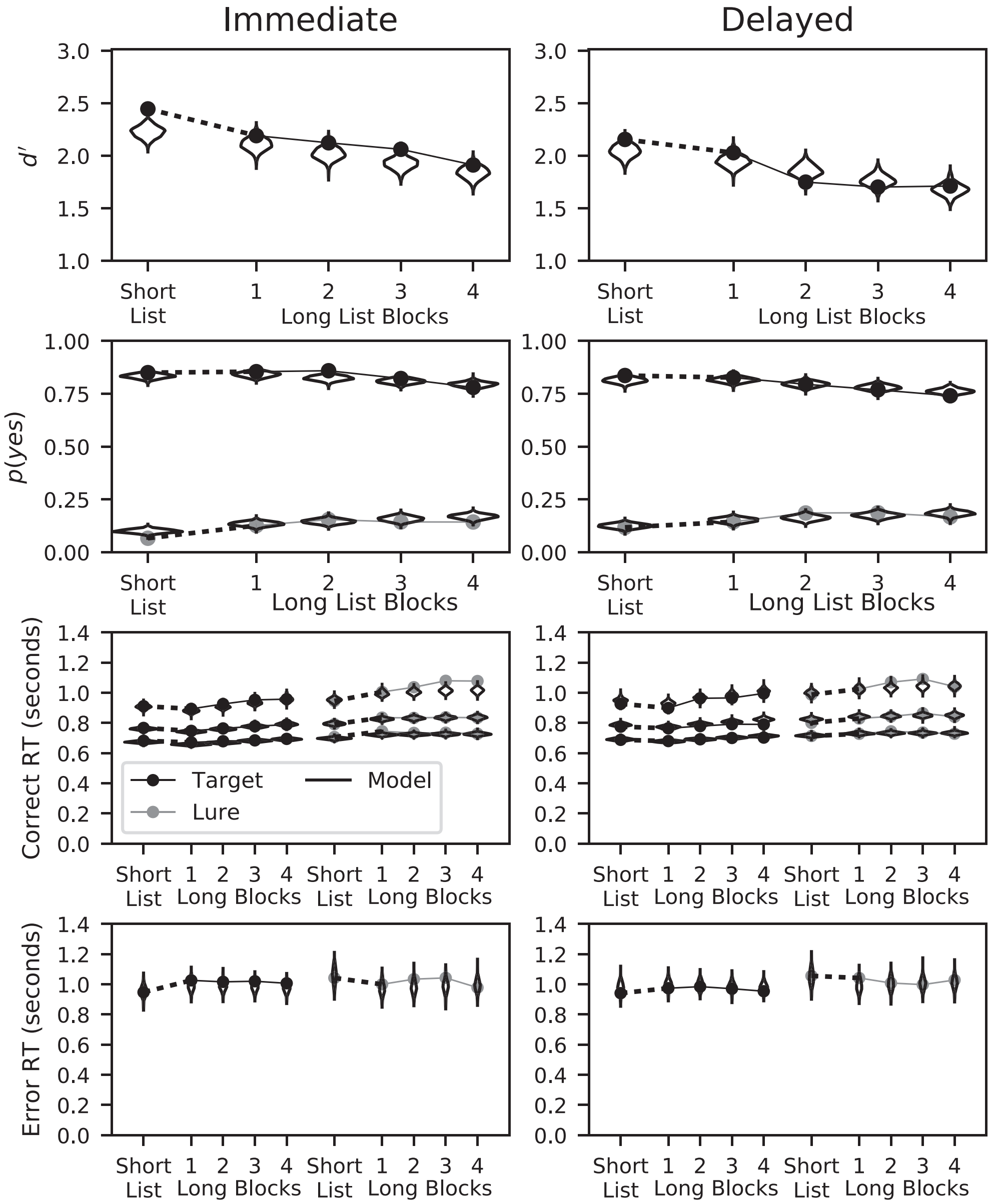

Fig. 5. Experiment 1 data and posterior predictive distributions for the immediate (left column) and delayed (right column) conditions. Depicted are d' (top row), choice probabilities (second row), correct RTs (third row), and error RTs (bottom row) for the short list and the long list, which was divided into four blocks of 24 trials. RT distributions are summarized using the average of each participant's 10^{th},50^{th}, and 90^{th} RT distribution percentiles for correct responses and the 25^{th},50^{th}, and 75^{th} percentiles for error responses.

increasing bias to say “old” (Gehring, Toglia, & Kimble, 1976), and thus we additionally allowed the starting point $z/a_{int.}$ to vary across the immediate and delayed conditions. All model parameters and the experimental factors they vary across can be found in Table 3.

8. Results and posterior predictives

Posterior predictive values were used to evaluate how well the model matches the data. Space prohibits depiction of all 107 participants' plots, and thus we focus on the average across participants, although plots of some of the individual participants' data and posterior predictive distributions can be found in the Supplementary Materials, and it was generally found that the model gave a strong account of the data from individual participants.

Fig. 5 depicts data and predictions for the short list along with the long list, which was broken up into four blocks of 24 trials. The figure depicts choice probabilities (top row) along with correct and error RTs (middle and bottom rows) for the immediate and delayed conditions (left and right columns) collapsed across the two WF conditions and averaged over participants. Due to the small number of trials in each cell, and the infrequency of errors ($M = 2.06$ errors per block), correct RTs are summarized using the mean of each participant's 25^{th},50^{th}, and 75^{th} percentiles for correct RTs, while the error RTs are summarized using the mean of each participant's median error RT.

One can see that the model is reproducing several of the trends in the data. Consistent with previous work, the list length effect, as measured by the difference in performance between the short list and the first block of the long list, was quite small but somewhat larger for the immediate ($\Delta d' = .26$) than the delayed condition ($\Delta d' = .13$, Cary & Reder, 2003; Dennis et al., 2008). The model captures the small list length effect in the delayed condition but falls short of capturing the larger effect in the immediate condition. The model naturally predicts smaller list length effects in delayed conditions because the magnitude of the item noise contribution depends on the match to the study context, which is impaired in the delayed condition. However, the magnitude of the delay effect was not sufficient to similarly predict a much smaller list length effect in the delayed condition. In previous applications this was accomplished by assuming poor context reinstatement in the long list of the immediate condition (Turner, Dennis, & Van Zandt, 2013; Osth & Dennis, 2015). We initially implemented this by assuming a separate context match parameter ξ for the long list in the immediate condition in addition to the ξ_{del} parameter for the delayed condition, but such a model demonstrated poor parameter recovery compared to a more restricted model where ξ only varies across conditions.

Similar to findings from Criss et al. (2011), the TPE is most evident as a decline in the HR over the test blocks (Δ HR $= -.08$ from the first to the last block, in both conditions), although there are slight increases in the FAR as well (Δ FAR $= .02$ in both conditions). The TPE is accompanied by slower correct RTs through the course of testing (previously reported in Murdock & Anderson, 1975), most evident in the later percentiles of the RT distribution. The error RTs, in contrast, are affected very little by the course of testing. The model captures all of these trends.

It should be noted that the changes in RT over testing are quite consistent with the predictions of item noise and context drift in Fig. 3 but are not congruent with decreases in speed-accuracy thresholds through the course of testing, as such decreases entail decreases in RT for both correct and error responses and are noticeably very large for error responses. Nonetheless, the model is sufficiently complex that the combined effects of the model parameters can operate in a way that defies intuition and requires formal analysis. We demonstrate below that the changes in speed-accuracy threshold are unlikely to drive the TPE as the mean change over participants did not differ significantly from zero. However, we found there was substantial variability across participants that moderates the size of the effect, with some decreasing their speed-accuracy thresholds over the course of testing while others increased them.

Posterior predictive plots for the marginal effects of each of the manipulations, namely word frequency, list length, and study-test delay, can be found in the Supplementary Materials. These plots demonstrate that the model is capable of addressing both the choice probabilities and RT distributions for each of these effects.

A reviewer inquired as to whether targets perform better when preceded by a target on the test list. HRs for targets preceded by targets ($M = .824$) were higher than when preceded by lures ($M = .793$). Although the effect is small, it is highly reliable, $BF_{10} = 695$. However, because items were tested in the same order as they were studied, it is unclear whether this was a sequential effect (Malmberg & Annis, 2012) or priming from an adjacent target on the study list (McKoon & Ratcliff, 1979). The same effect was present for lures, in that FAR were higher for lures preceded by targets ($M = .161$) than for lures preceded by lures ($M = .127$), $BF_{10} = 9758$, making it likely that sequential effects accounted for part of the effect for targets. A thorough analysis by Schwartz, Howard, Jing, and Kahana (2005) found facilitation when targets were tested by adjacent targets versus non-adjacent targets, but the effect was only found for high confidence responses.

9. Analysis of parameter estimates

The posterior predictives revealed that the model is able to reproduce the key trends in the data. With a complex non-linear models such as ours, it can be difficult to ascertain which parameter, or combination of parameters, is responsible for the observed effects by analyzing the parameters alone. Instead, we generated the predicted TPE for each posterior sample and correlated it with their parameters to ascertain which parameter was responsible for the decline. In particular, we generated a predicted discriminability decrement, d'_{change} by calculating the predicted d' for the first trial (d'_1) and subtracting it from the predicted d' for the last trial of the long list d'_{96} based on 10,000 simulations for a target and a lure in the first and last trial in each condition. This was done for 20% of posterior samples for each participant.

We calculated correlations between d'_{change} and the three major culprits of the TPE, namely the context drift parameter γ, the item mismatch variability parameter σ^2_{ti}, which governs the total degree of item noise, and the total change in response boundaries over the test sequence a_{change} (where a_{change} is a_{slope} multiplied by the number of test trials in the long list). To measure uncertainty in the correlation, we performed a bootstrap analysis where posterior parameters were sampled randomly with replacement; the correlation was calculated on each bootstrap iteration and the procedure was repeated 1000 times. Scatterplots along with 2D kernel density estimates of each of these comparisons for the immediate and delayed conditions can be seen in Fig. 6. 95% highest density intervals (HDIs) for the correlation are reported in brackets.

Both the immediate and delayed conditions display largely similar results. Recall that when $\gamma = 1$, context is completely preserved from trial-to-trial and no context drift occurs, whereas when $\gamma < 1$, context drifts and forgetting occurs. Thus, any extent to which γ predicts forgetting through the test list will be reflected in positive correlations. Indeed, γ shows large correlations in each condition ($r \sim .68$). σ^2_{ti} governs the total amount of item noise, and should cause greater forgetting through the test list as it is increased. Thus, any extent to which σ^2_{ti} is predictive of forgetting through the test list should be reflected in a negative correlation between σ^2_{ti} and

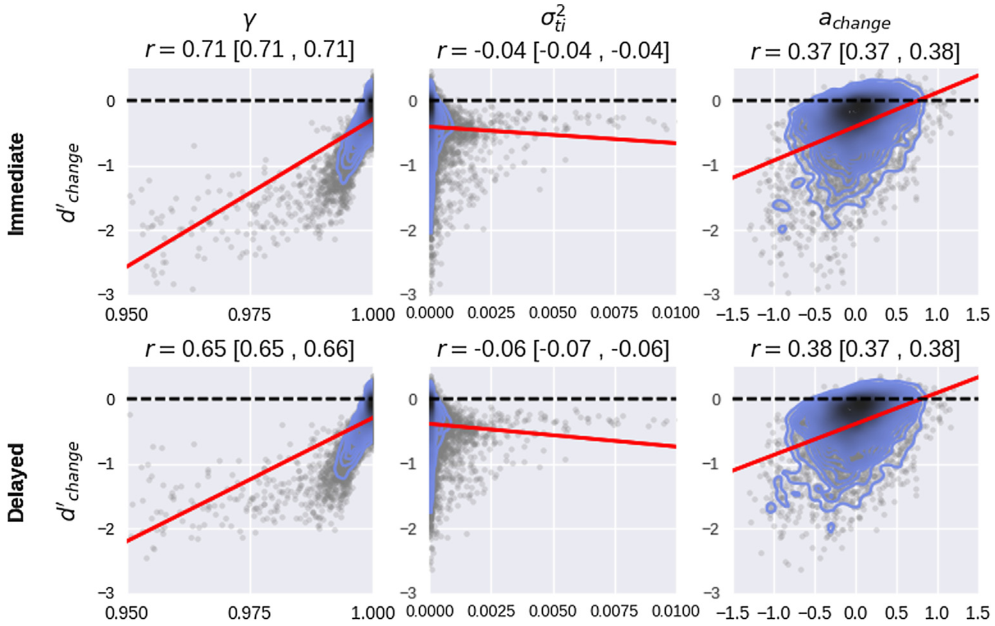

Fig. 6. Scatterplots (grey dots), 2D density estimates (shaded regions), and correlation coefficients (with upper and lower bounds of the 95% HDI reported in brackets) of the predicted d' decline through the long test list against the parameters γ,σ^2_{ti}, and a_{change} for the immediate (top row) and delayed (bottom row) conditions. Linear regression lines are shown in red. Darker areas of the density plot depict the areas with more numerous posterior samples.

d'_{change}. Interestingly, σ^2_{ti} showed only small negative correlations in each condition ($r \sim -.05$). a_{change}, which reflects the total degree of change in the speed-accuracy threshold, should show a positive relationship with d'_{change} if it predicts testing related forgetting. Indeed, a_{change} showed substantial positive correlations in each condition ($r \sim .375$).

In the Supplementary Materials, we repeat the same analysis on a set of models that lack some of the candidate mechanisms (described in more detail in the next section). Each of the models that contained both context drift and item noise revealed a strong role of context drift in predicting the TPE and only a minor role of item noise, even when changes in response bias, changes in response boundaries, or both mechanisms were removed from the model.

To get a sense of the inter-participant variability in the changes and response bias and boundaries, the median of each participant's posterior slopes was multiplied by 96 (the number of test trials in the long list) to produce a measure of the total magnitude of change over the test. These results can be seen in the left column of Fig. 7. Similar to the results of Ratcliff (1978), participants become more conservative (i.e., less likely to classify a test item as studied), as evidenced by a decreasing response bias z/a as testing

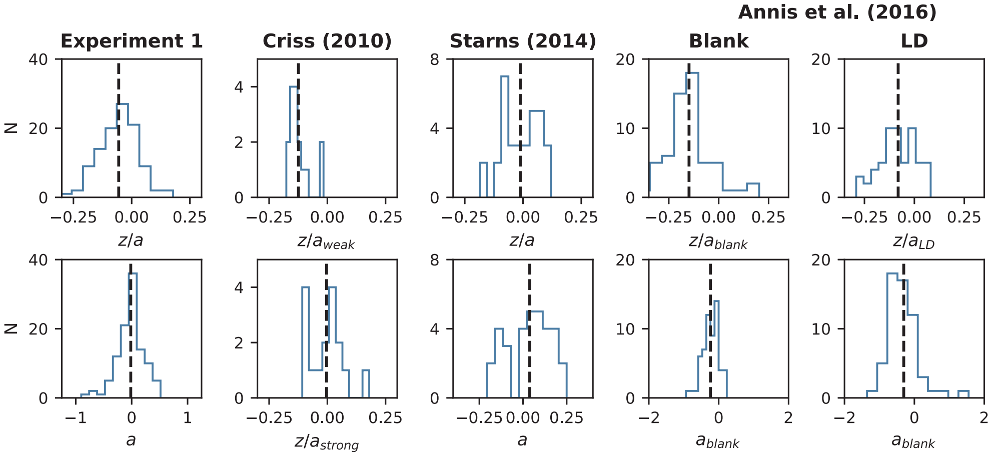

Fig. 7. Histograms of the median of each participants posterior distribution of the change in z/a and a over the long test list for Experiment 1 (left column), the Criss (2010) dataset (second column), the Starns (2014) dataset (third column), along with the blank (fourth column) and LD conditions (fifth column) of the Annis et al. (2016) dataset. Note that only the changes in z/a are shown for the Criss (2010) dataset because the $a_{slope} = 0$ model was the preferred model in the model selection procedure.

progressed. This was confirmed by an analysis of the group mean for z/a_{change}, which did not include zero in their 95% HDI, $M = -.055$ $[-.078, -.034]$ (95% HDI in brackets). Changes in response boundary, in contrast, did not show a consistent pattern across participants: the group mean of a_{change} did not deviate significantly from zero, $M = -.015$ $[-.088, .055]$. Nonetheless, Fig. 7 demonstrates that there was considerable variability in this slope across participants, with some participants increasing their speed-accuracy thresholds through the test and some decreasing them, which has a strong moderating role on the effect size of the TPE.

In the Supplementary Materials, we present the results of a parameter removal analysis as an alternative to the correlation analysis described above. The analysis was performed with the same goal in mind – to better understand which culprits are allowing for the prediction of the TPE in the full model. In this analysis, we generated the predicted TPE from the full model as described above. Subsequently, we generated the same predictions, but where only one culprit was allowed to apply. That is, we generated predictions where (a) only context drift was allowed to operate through the course of testing (item noise and response boundaries were fixed across test trials), only increases in item noise were allowed to operate (context and response boundaries were fixed), and only changes in response boundaries were allowed (context and item noise were fixed). Context drift produced decreases in performance through testing that were closest to the full model, while item noise produced only small decreases in performance and changes in response boundary produced a mean change in performance that was close to zero. This alternative analysis supports the conclusion that context drift is the largest predictor of the TPE.

While the results of Fig. 6 suggest that item noise is not responsible for the TPE, this begs the question – do the data necessitate item noise? While Fig. 5 showed a small effect of list length on d', it's possible that the effect is due to lower speed-accuracy thresholds in the long list condition. To investigate this, we performed an analysis where we compared the predicted list length effect, $d'_{long}-d'_{short}$, to the item mismatch variability parameter σ_{ti}^2 along with the difference in response boundaries between the short and long list conditions, $a_{long}-a_{short}$. To avoid the confounding effect of test position, long list predictions were generated for the first 24 trials only, which is equal to the number of short list trials. The results can be found in Fig. 8. Relatively large correlations were found between σ_{ti}^2 and the predicted list length effect ($r \sim -.485$). Thus, unlike previous results in short-term recognition memory which found that list length effects could be accounted for entirely by changes in speed-accuracy thresholds (Donkin & Nosofsky, 2012b), the present results suggest that there are changes in memory strength due to increasing item noise. This result is reinforced by the model selection analysis in the next section, which found there were penalties for removing item noise from the full model.

The difference in speed-accuracy thresholds across the two list length conditions also exhibited large correlations with the predicted list length effect, with values of .53 and .55 in the immediate and delayed conditions, respectively. However, as was the case with the TPE, on average participants did not have a strong tendency to shift their response boundaries across list length, as evidenced by an analysis of the difference in group mean of $a_{int.}$ between the short and long list conditions, which included zero in the 95% HDI: $M = -.04$ $[-.14, .07]$.

9.1. Model selection

In the previous section, we measured the extent to which each model component was responsible for the TPE. Here, we ask a

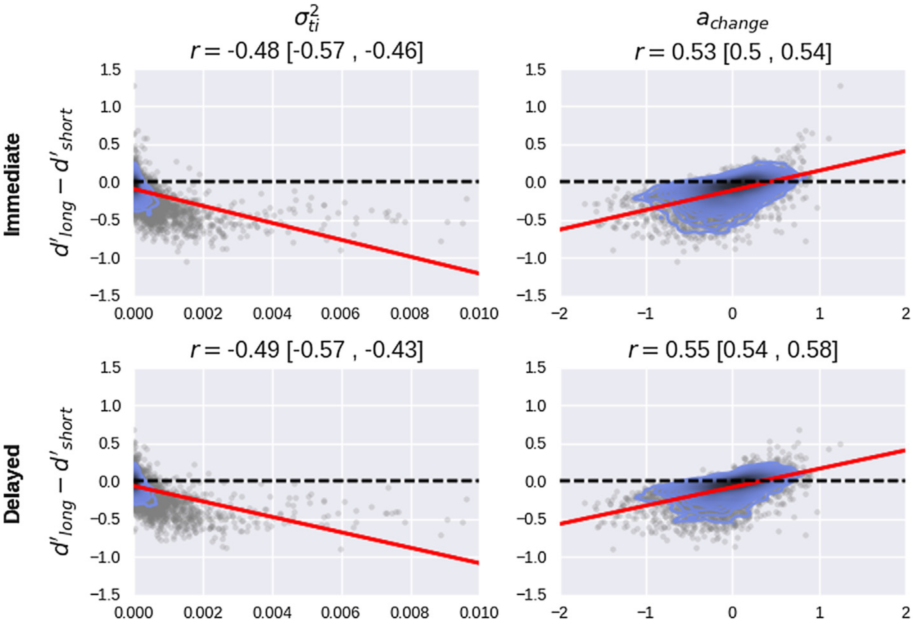

Fig. 8. Scatterplots (grey dots), 2D density estimates, and correlation coefficients (with upper and lower bounds of the 95% HDI reported in brackets) of the predicted list length effect $d'_{long}-d'_{short}$ against σ_{ti}^2 and $a_{long}-a_{short}$ for the immediate (top row) and delayed (bottom row) conditions. Darker areas of the density plot depict the areas with more numerous posterior samples.

Table 4
Δ WAIC values for the restricted models for each dataset.

	Exp. 1	Criss (2010, E2)	Starns (2014, E2)	Annis et al. (2016, E2)
$\sigma_{ti}^2 = 0$	−25	−5	−57	10
$\gamma = 1$	−50	−110	−32	−9
$z/a_{slope} = 0$	−120	4	−64	−345
$a_{slope} = 0$	−31	61	−8	−282
$z/a_{slope}, a_{slope} = 0$	−286	30	−45	−716

Notes: σ_{ti}^2: item mismatch variability parameter, γ: context drift parameter, Δ WAIC: WAIC difference between the restricted model and the full model, with positive values indicating improvement over the full model.

stronger version of the question by removing each component of the model and evaluating whether or not a simpler model produces an advantage over the full model that contains all three components. To do this, we used model selection techniques which quantify the complexity of the model and subtract that from a measure of its ability to fit the data. We employ the widely applicable information criterion (WAIC: Watanabe, 2010) for this purpose. Smaller values of WAIC mean that a model strikes a better balance between goodness-of-fit and simplicity.

We fit a total of five additional models to the data, each of which fixes one or more parameters to eliminate estimation of one of the components: a model with the item mismatch variability parameter set to zero ($\sigma_{ti}^2 = 0$), which completely lacks item noise, a model with no context drift ($\gamma = 1$), a model with no changes in response bias ($z/a_{slope} = 0$), a model with no changes in the speed/accuracy threshold ($a_{slope} = 0$), along with a model where both bias and speed-accuracy threshold are both fixed across test trials (z/a_{slope} and $a_{slope} = 0$), meaning that all changes in performance through testing are memory related. Because WAIC is only meaningful when compared to another model, we calculated Δ WAIC for each model by subtracting each model's WAIC score from the full model. Positive values indicate improvements relative to the full model. Because WAIC is measured on a log likelihood scale, Δ WAIC scores greater than 10 are conventionally considered large. Δ WAIC values for each model for all of the datasets in this article can be seen in Table 4.

Interestingly, in Experiment 1, none of the models outperformed the full model. While the weak correlations between item noise and the magnitude of the TPE suggest that item noise is not likely to be responsible for the TPE, the $\sigma_{ti}^2 = 0$ model still exhibits substantial penalties for its lack of item noise. Nonetheless, the Δ WAIC decrement for the σ_{ti}^2 was the smallest of the restricted models. Inspection of the posterior predictives (not depicted here for space considerations) of the $\sigma_{ti}^2 = 0$ model showed that it was not able account for the list length effects in the data as well as the full model, which provides further evidence that changes in speed-accuracy thresholds across the list length conditions is not a sufficient account of the list length effect in recognition memory. In contrast, the $\gamma = 1$ model, which lacked context drift, was capable of addressing the list length effects seen in the data but was not able to predict a steep enough decline through the course of testing.

9.2. Discussion

We applied the model to an experiment that manipulated list length, study-test delay, and word frequency. Each manipulation constrained a relevant model parameter, with both list length and test position constraining the item noise parameter, study-test delay constraining the intercept of the contextual drift function, and word frequency constraining the context-noise parameters. Not only was the model able to provide a very good account of the data, the resulting parameter estimates suggested that item noise plays only a small role in predicting the TPE. This was likely because the list length manipulation produced only a small effect on performance, consistent with previous investigations. This is rather constraining on predictions for the TPE because as shown in Fig. 3, the accrual of item noise has a diminishing impact on performance as more items are added to memory. Consequently item noise predicts that the list length effect must be substantially *larger* than the detriment of recognition testing. Nonetheless, item noise was clearly necessary to account for the list length effects present in the data, as evidenced by the fact that the model with no item noise (the $\sigma_{ti}^2 = 0$ model) did not perform as well as the full model. Thus, the results of our modeling do not support the conclusions of models which completely lack item noise (e.g.; Dennis & Humphreys, 2001).

The context drift process, in contrast, appeared to be most predictive of the TPE in both conditions. As shown in Fig. 3, with only a rather slight degree of contextual drift, over the course of 96 trials a rather substantial degree of forgetting occurs that is specific to the test period. A surprising result was that changes in response boundaries were also found to be quite predictive of the TPE, despite the fact that there was no consistent trend across individuals in their tendency to change response boundaries through the course of testing. In other words, some participants demonstrated large increases in response boundaries through testing, improving performance and mitigating against increasing item noise and contextual change, while others demonstrated large decreases in response boundaries through testing, which further exacerbate the detrimental effects of recognition testing.

In the coming sections, we evaluate the generalizability of a contextual drift account by fitting additional datasets which have been used to argue for an item noise account of the TPE.

10. Pure strength lists: Criss (2010, Experiment 2)

While the modeling results of Experiment 1 argue that the test position effect is mostly due to contextual change and not due to item interference, a recent set of results manipulating the strength composition of study lists has been used to argue for an item noise account of the TPE where a differentiation process is occurring through the course of testing. Kiliç et al. (2017) found that when participants studied pure strength lists, such as a pure weak list where all items were encoded shallowly and a pure strong list where all items were encoded deeply, HRs declined at a steeper rate for items in weak lists than for items in strong lists. However, when participants studied a mixed strength study list (50% weak items and 50% strong items), HRs declined at roughly the same rate for weak and strong items.

Kiliç et al. (2017) demonstrated that their variant of the REM model which employs a differentiation process during testing predicted these qualitative trends in pure and mixed strength lists. In differentiation models, repetitions of items accumulate into a single strong trace that responds strongly to its own cue and weakly to other cues (Criss, 2006; Shiffrin, Ratcliff, & Clark, 1990). Functionally, differentiation produces a reduction in interference from studied items as strength is increased. In the Criss et al. (2011) variant of the TPE, differentiation occurs during the process of testing, but only for tested items that are recognized. Subsequently, the trace with the strongest likelihood ratio (which is most likely to correspond to the target) is selected to be updated with additional features and this reduces item noise. Test items that are not recognized do not produce differentiation. Instead, a new trace is added to memory, which increases interference on subsequent test trials.

In pure strong lists, target items are likely to be recognized and undergo differentiation. On subsequent test trials, the memory traces corresponding to previously tested items that underwent differentiation will exert less interference on the currently tested item, which has the effect of reducing the TPE. In pure weak lists, in contrast, targets are less frequently recognized due to their weak encoding. These unrecognized items create new memory traces, increasing interference on later test trials and producing a steeper TPE. Mixed strength lists demonstrate both influences, with weak items increasing interference and strong items decreasing it, resulting in roughly equal declines in performance for both weak and strong items.

Can a contextual drift account address the list strength results described by Kiliç et al. (2017)? If our model is unable to address such results, it would provide a strong case for their item noise account. Interestingly, in our model context drift in combination with the changes in response bias appear to make the similar predictions as the REM model for both pure and mixed strength lists despite our model lacking a differentiation process. To demonstrate this, we averaged all of the participant parameters from the immediate condition of the previous dataset for the $\sigma_{ti}^2 = 0$ model (which lacks item noise) and simulated performance for the experiment. To simulate performance in the weak and strong conditions, the learning rate r was set to .5 and 2.0 for each respective condition. For the mixed strength case, expected log likelihood ratio distributions were generated using $r = 1.25$ (the average of the weak and strong learning rates) and the likelihood ratio transformation was calculated according to Eqs. (19) and (20) (present in Appendix A, along with details of the log likelihood ratio transformation).

HR and FAR predictions from the model for each test trial in pure and mixed strength lists can be seen in Fig. 9. For the pure strength lists, HR declines at a much steeper rate for weak (Δ HR = −.09 over the course of testing) than for strong items (Δ HR = −.04). For the mixed strength lists, the HR declines are more similar for weak (Δ HR = −.05) and strong items (Δ HR = −.03). In this simulation, the only critical difference between these two conditions concerns the nature of the expected strengths in the likelihood ratio transformation (see Appendix A for details). In pure lists, weak items are expected to be weak and strong items are expected to be strong, while in mixed lists, both weak and strong items are expected to have an intermediate

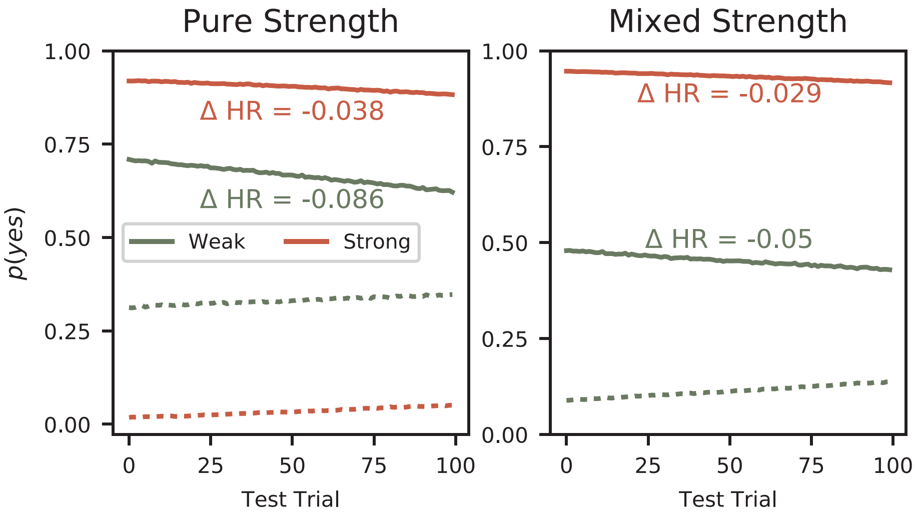

Fig. 9. Simulated predictions from the model for pure (left) and mixed strength (right) lists. Parameters of each model were estimated from the list length dataset, with the exception of r_{weak} and r_{strong} which were set to .5 and 2.0, respectively, which are as follows: $z/a_{int.} = .58, z/a_{slope} = -.000587, a_{int.} = 1.68, a_{slope} = -.000130, d_c = .03, \sigma_{ti}^2 = .174, \rho = .065$, and $\gamma = .997$.

strength. As a consequence, the mean of the target distribution is higher for weak items in the mixed list relative to the pure weak list, while the opposite is the case for strong items, which have a lower target distribution mean in the mixed list relative to the pure strong list. Note that these predictions only hold when z/a_{slope} is negative; when z/a_{slope} is set to zero while all other parameters are kept the same, the difference in HR between strong and weak items is quite small. Nonetheless, as we will demonstrate throughout the article, the finding that participants become increasingly conservative throughout the duration of the test list is relatively common and also replicates the findings of Ratcliff (1978).

A potential weakness of the work of Kiliç et al. (2017) is that all of the differences in their observation of the TPE across test lists must necessarily be attributed to memory retrieval. Because they did not model response times, they were not able to ascertain whether the differences in conditions also affected speed-accuracy thresholds. Inferences about pure list manipulations of strength are complicated by the fact that participants may adopt different biases and speed-accuracy thresholds in each condition. Indeed, analyses of response times suggest that participants employ a starting point closer to the “yes” boundary in conditions of higher list strength, despite exhibiting a reduced FAR in such conditions (Criss, 2010; Kiliç & Öztekin, 2014; Starns, Ratcliff, & White, 2012). Our model circumvents the limitations of previous work by allowing for the simultaneous investigation of factors related to memory retrieval and decision making.

In order to directly test the effect of a pure list strength manipulation on the TPE we fit data from Criss (2010, Experiment 2). In this dataset, words were presented either once or five times, and both HF and LF words were used. Both strength and word frequency were manipulated between list. Because this dataset did not contain a list length manipulation, we borrowed constraint from Experiment 1 by using Experiment 1’s posterior distributions as informed priors for the fit to this dataset. This was done for the memory retrieval parameters only, as decision parameters, such as response bias, response boundaries, and nondecision time, can vary depending on instructions, presentation format, or perceived difficulty of the stimuli. Informed priors were constructed using kernel density estimation (KDE) for the posterior distributions for the group parameters of the context mismatch variability parameter ρ, the item match variability parameter σ_{tt}^2, the item mismatch variability parameter σ_{ti}^2, and the context drift parameter γ. The complete list of model parameters for this dataset, along with which parameters were constrained by informed priors, can be seen in Table 3.

10.1. Parameterization

To account for the effects of the strength manipulation, learning rate parameters r_{weak} and r_{strong} were introduced to account for the different effects of presentation rate. To impose additional constraint on the model, we fixed the average context noise, ρ_{avg}, to .039, which was the mean of the group mean distribution’s estimate for Experiment 1. Different linear functions of response bias and boundaries were allowed for the weak and strong conditions, in part due to the fact that strength was manipulated cross-list but also because of the finding that list strength manipulations affect the bias parameters of the DDM. Additionally, because testing occurred a relatively short time after the study list was completed, the context reinstatement parameter ξ was fixed to one. The complete list of parameters can be found in Table 3.

10.2. Posterior predictives

The model selection results in Table 4 indicated that the $a_{slope} = 0$ model was the preferred model for this dataset by a substantial margin. For that reason, all analyses here are restricted to the $a_{slope} = 0$ model instead of the full model. Fig. 10 depicts data and predictions for the four conditions of the experiment. Test lists were grouped into four blocks of 25 trials. The figure depicts choice probabilities (top) along with correct and error RTs (middle and bottom panels) averaged over participants. Due to the infrequency of errors ($M = 9.41$ errors per cell), each participant’s error RTs were summarized using the .25, .5, and .75 quantiles of the RT distribution, whereas correct responses were summarized using the .1, .5, and .9 quantiles.

The figure shows that the model is providing a good account of the data. The TPE is primarily manifested as a decrease in HR over trials, with a steeper decline in the weak conditions (Δ HF HR = −.21, Δ LF HR = −.19, all differences represent the differences between the first and the last block) relative to the strong conditions (Δ HF HR = −.05, Δ LF HR = −.09). The model reproduces this trend, and although it underpredicts the magnitude of HR decline in the weak conditions (MΔ HF HR = −.16, MΔ LF HR = −.16) and slightly overpredicts the magnitude of decline in the strong HF condition (MΔ HF HR = −.09, MΔ LF HR = −.08), it otherwise reproduces the qualitative pattern. FARs showed increases in all conditions that ranged from slight to moderate in magnitude (Δ HF weak FAR = .04, Δ LF weak FAR = .05, Δ HF strong FAR = .07, Δ LF strong FAR = .01).

Correct RTs increased over the course of testing, while error RT was largely unaffected. Each of these patterns was reproduced by the model. The RT trends are again consistent with the item noise and context drift predictions depicted in Fig. 3 but not with decreases in speed-accuracy thresholds, which predict substantial decreases in RT that are larger for error than correct responses.

Analyses and plots of individual participants’ data can be found in the Supplementary Materials, where it is demonstrated that the model gave a reasonable account of these data as well.

10.3. Analysis of parameter estimates

We have demonstrated that our model is capable of predicting a steeper decline in performance in pure weak tests than in pure strong tests. However, a question remains as to whether or not the observed declines are being driven by context drift or by increasing item noise. Thus, here we reproduced the parameter correlation analyses for the Criss (2010) dataset to evaluate whether or not similar conclusions can be reached from these data. Because the winning model in Table 4 lacked changes in speed-accuracy

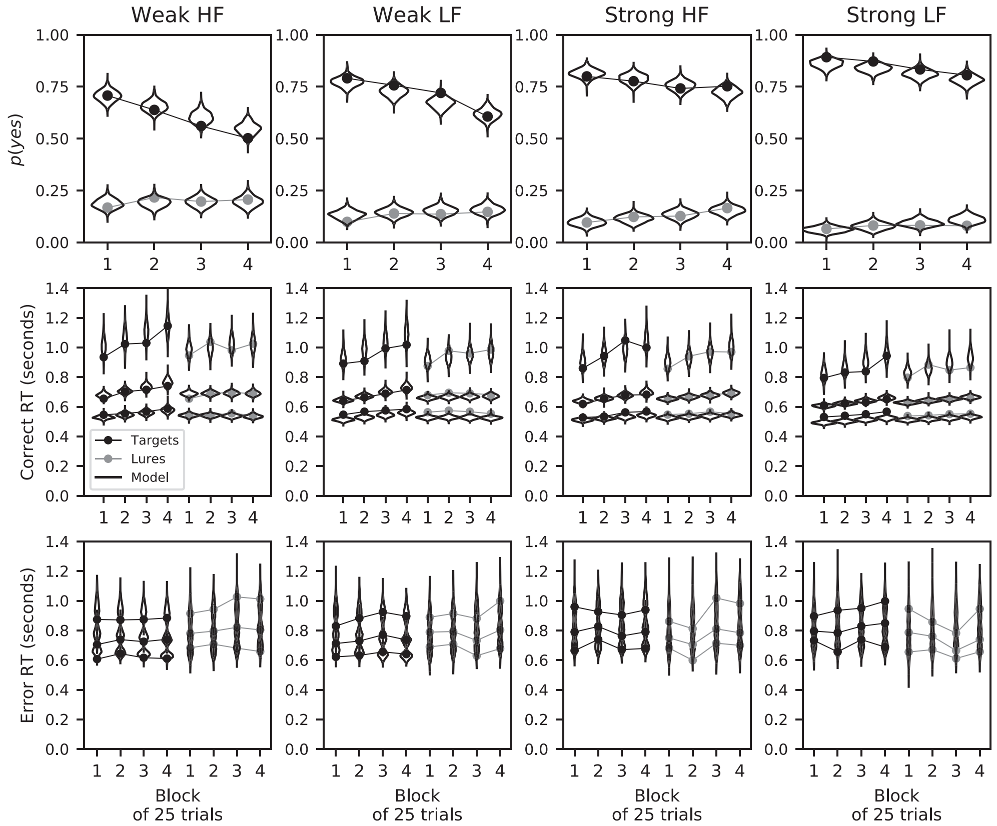

Fig. 10. Data and posterior predictive distributions for the data from Criss (2010, Experiment 2) for each test block (25 trials per block) of the four conditions. Depicted are the choice probabilities (top) along with the correct (middle) and error (bottom) RTs for both targets (T) and lures (L). Correct RTs are summarized using the mean of each participant's .1, .5, and .9 quantiles of the RT distribution, while the error RTs are summarized using the .25, .5, and .75 quantiles.

thresholds through the course of testing, correlations between d'_{change} were restricted to the other culprits of the TPE, namely the context drift parameter γ and the item mismatch variability parameter σ^2_{ti} which governs the total degree of item noise. Scatterplots along with 2D kernel density estimates of each of these comparisons for each experimental condition can be seen in Fig. 11.

Analyses reproduced the key findings from fits to Experiment 1. The context drift parameter γ had the strongest correlation with the TPE, ranging from .49 to .73 across each condition. The item mismatch variability parameter, σ^2_{ti}, showed moderately positive correlations in each condition, implying that higher item noise is associated with *less* of a decline across testing. These analyses converge with prior analyses suggesting that item noise appears to play only a minor role in producing the TPE.

Analyses of the full model (which includes changes in speed-accuracy thresholds through the course of testing) along with the other alternate models in Table 4 can be found in the Supplementary Materials; these analyses reinforce the conclusions above.

To get a sense of the inter-participant variability in the changes and response bias, the median of each participant's slope was multiplied by 100 (the number of test trials) to produce a measure of the total magnitude of change over the test. Histograms of these parameters can be seen in the second column of Fig. 7. Inspection of the figure reveals that more participants become increasingly conservative through the course of testing in the weak conditions, as measured by negative values of z/a_{change}. For the strong conditions, in contrast, tendencies to become more conservative and liberal were more evenly distributed. This difference was reflected in the group mean parameters for each condition. The 95% HDI for z/a_{change} for the weak condition did not include zero ($M = -.12$ $[-.16, -.08]$), whereas zero was included in the HDIs for the strong condition ($M = 0$ $[-.05, .05]$). These results suggest that one contribution to the steeper decline in HR in the weak conditions is a greater tendency for participants to become more conservative throughout the test.

10.4. Discussion

In line with the predictions in Fig. 9, the model was able to reproduce the greater decline of HR in pure strong lists relative to pure weak lists. Analyses of the posterior predictives demonstrated that context drift was the primary culprit for the TPE in this dataset. Nonetheless, the data from Criss (2010, Experiment 2) was restricted to pure lists of either weak or strong items. To test the predictions for mixed strength study lists, we employed data from Starns (2014, Experiment 2).

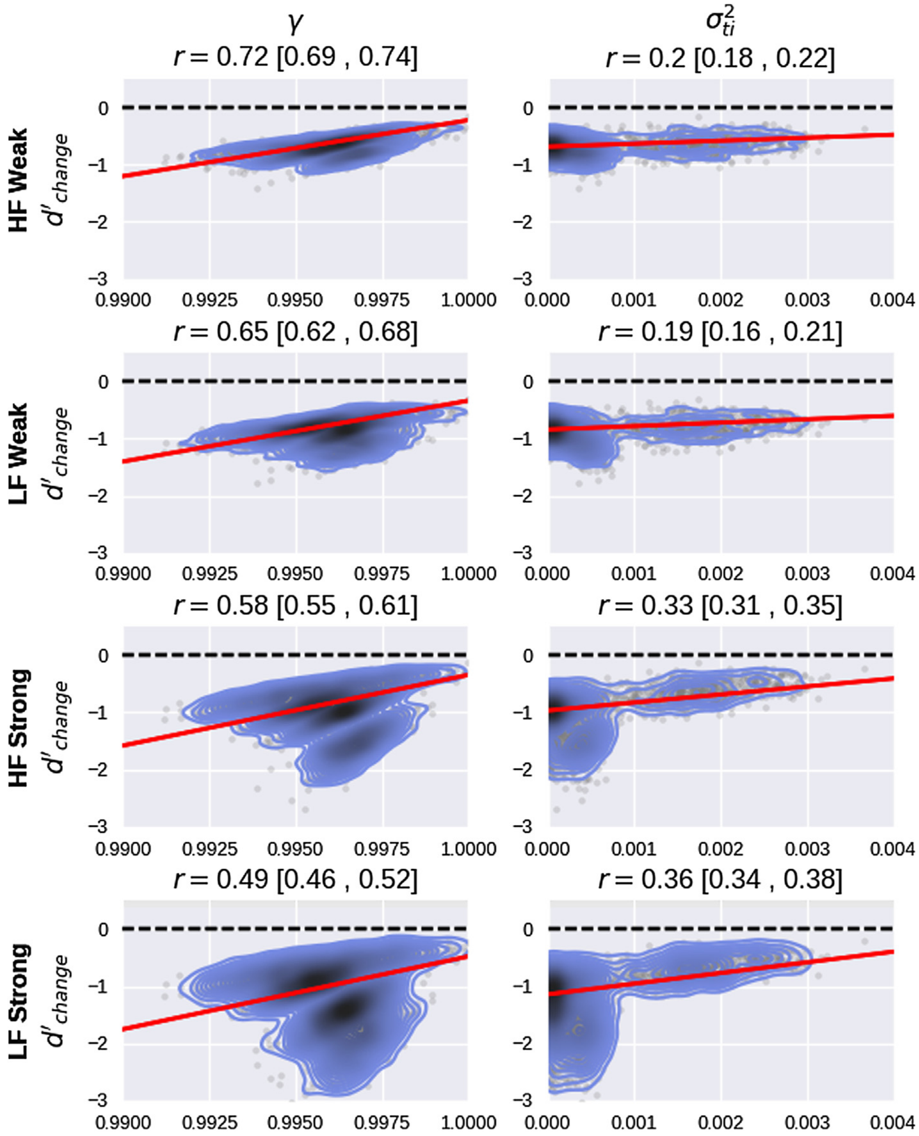

Fig. 11. Scatterplots (grey dots), 2D density estimates (shaded regions), and correlation coefficients (with upper and lower bounds of the 95% HDI reported in brackets) of the predicted d' decline through the test list against the parameters γ and σ^2_{ti} for each condition of the Criss (2010, Experiment 2) dataset. Results are depicted for the $a_{slope} = 0$ model, which was the preferred model in the model selection procedure. Regression lines are shown in red. Darker areas of the colored density depict the areas with the highest posterior density. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

11. Mixed strength lists: applying the model to data from Starns (2014, Experiment 2)

Our model predicts that the HR decline with test position should be similar for weak and strong items in mixed lists of strong and weak items. To test this prediction, we applied the model to data from Starns (2014, Experiment 2). In this dataset, words were studied either once or three times all within a single study list. To inherit the constraint from the prior datasets, we used the posteriors from the fit to Criss (2010, Experiment 2) as informed priors for this dataset for a number of the memory model parameters.

11.1. Parameterization

Learning rate parameters r_{weak} and r_{strong} were used to account for the different effects of presentation rate. Because the presentation rates were different from the Criss dataset, the learning rate parameters used relatively uninformative priors. In addition, because words appeared to be of medium frequency in this dataset, we fixed ρ to .039, which was the same value as ρ_{avg} in the application to the previous dataset. Given that only a single study list contained all the different conditions of the experiment, the linear functions on a and z/a did not vary across any conditions. Similar to the Criss dataset, a relatively short retention interval was

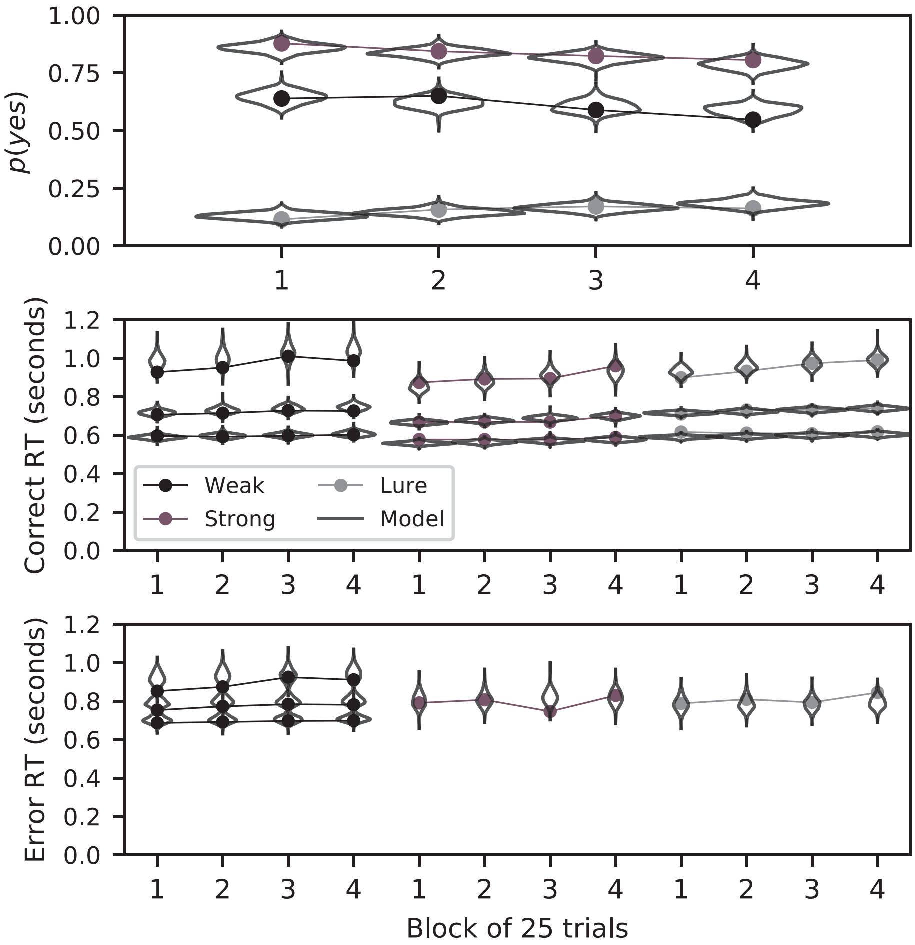

Fig. 12. Data and posterior predictive distributions for the data from Starns (2014, Experiment 2) for each test block (25 trials per block) for the weak (W), strong (S), and lure (L) items. Depicted are the choice probabilities (top) along with the correct (middle) and error (bottom) RTs. Correct RTs are summarized using the mean of each participant's .1, .5, and .9 quantiles of the RT distribution, while the error RTs are summarized using the .25, .5, and .75 quantiles for weak items and the median RT for the strong condition and lures.

used in this experiment and thus the context reinstatement parameter ξ was fixed to one. The complete list of parameters can be found in Table 3.

11.2. Posterior predictives

Fig. 12 depicts data and predictions for the experiment. Test lists were grouped into four blocks of 25 trials. The figure depicts choice probabilities (top) along with correct and error RTs (middle and bottom panels) averaged over participants. Due to the infrequency of errors for strong items ($M = 2.98$) and lures ($M = 5.46$) in each cell, each participant's error RTs were summarized using the median RT. There were somewhat more errors to weak items ($M = 7.04$) in each cell, allowing the error RTs to be summarized using the .25, .5, and .75 quantiles. Correct responses were summarized using the .1, .5, and .9 quantiles.

Once again, the model provides an excellent account of both the choice probabilities and RTs. In contrast to the pure strength manipulations in the Criss dataset, this dataset shows very similar declines in HR across the two strength conditions (Δ weak HR $= -.09$, Δ strong HR $= -.07$; all differences are between the first and the last block). The model produces virtually identical declines across the weak and strong conditions ($M\Delta$ weak HR $= -.06$, $M\Delta$ strong HR $= -.07$). In addition, there was a moderate increase in the FAR in the data (Δ FAR $= .05$), and increasing correct RTs through the course of testing, whereas error RTs were not strongly affected by the course of testing. The model reproduces each of these qualitative trends. The RT trends are again consistent with the item noise and context drift predictions depicted in Fig. 3 but not with decreases in speed-accuracy thresholds, which predict substantial decreases in RT that are larger for error than correct responses.

Analyses and plots of the individual participants' data can be found in the Supplementary Materials, where it is found that the model provides a strong account of these data.

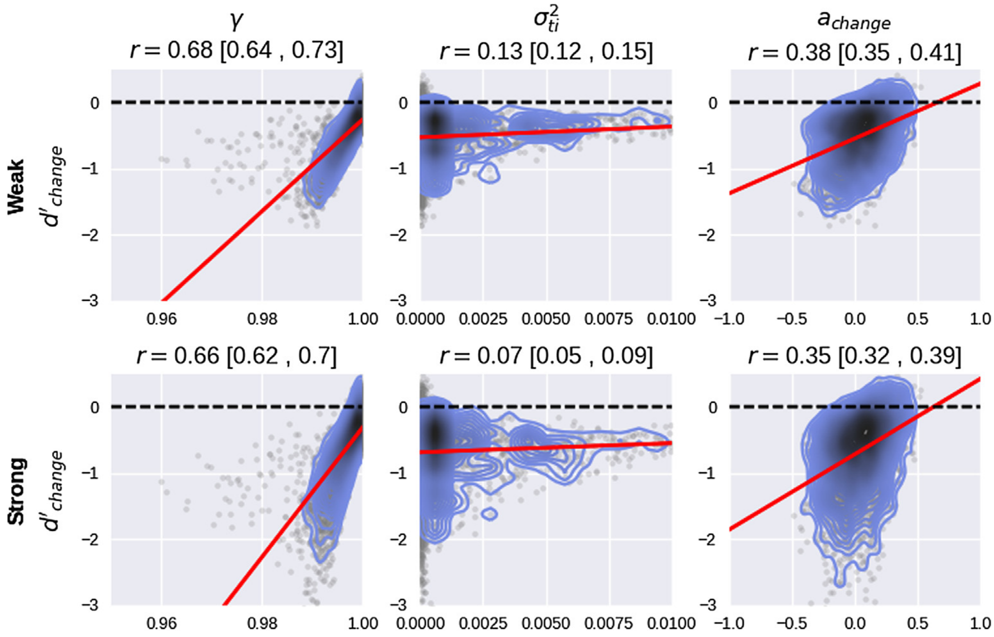

Fig. 13. Scatterplots (grey dots), 2D density estimates (shaded regions), and correlation coefficients (with upper and lower bounds of the 95% HDI reported in brackets) of the predicted d' decline through the test list against the parameters γ,σ^2_{ti}, and a_{change} for the weak (top) and strong (bottom) conditions of the Starns (2014, Experiment 2) dataset. Red lines are linear regression estimates. Darker areas of the colored density depict the areas with the highest posterior density. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

11.3. Analysis of parameter estimates

Correlations between d'_{change} and the three major culprits of the TPE, namely the context drift parameter γ, the item mismatch variability parameter σ^2_{ti} which governs the total degree of item noise, and the total change in response boundaries over the test sequence a_{change} (where $a_{change} = 100a_{slope}$), were calculated using the same procedure as before. Scatterplots along with 2D kernel density estimates of each of these comparisons for each experimental condition can be seen in Fig. 13.

Similar to previous analyses, the biggest predictor of the TPE was the context drift parameter γ, producing correlations ~ .7 for weak and strong items. Changes in the item mismatch variability parameter σ^2_{ti} again produced small positive correlations, implying that increasing item noise is associated with less of a decline in performance across trials. Changes in speed-accuracy thresholds were associated with relatively strong correlations for weak and strong items ($r \sim .36$).

Histograms of the changes in bias parameter can be found in the third column of Fig. 7. There was not a strong tendency to change bias over trials, as the 95% HDIs on the group means included zero for z/a_{change} ($M = -.01$ $[-.05, .03]$). The lack of increasing conservatism in this dataset replicates the lack of bias shift that occurred in the strong conditions of the Criss (2010) dataset. In addition, there was not a strong tendency for participants to change their speed-accuracy threshold over trials, as indicated by the fact that the 95% HDIs on the group means included zero for a_{change} ($M = .04$ $[-.03, .11]$). The lack of a clear bias shift along with the fact that the variability in a_{change} was relatively small compared to previous datasets may be part of the reason why the models that lacked these components (the $z_{slope} = 0$ and $a_{slope} = 0$ models) did not produce very large model selection penalties in Table 4 relative to the models that lacked memory related components (the $\gamma = 1$ and $\sigma^2_{ti} = 0$ models).

11.4. Discussion

In line with the predictions in Fig. 9, the model was able to reproduce the near equivalent declines in HR across the weak and strong conditions in the mixed list experiment of Starns's (2014) Experiment 2. The conjunction of this result along with the steeper decline in HR for weak items than strong items in pure lists in the data of Criss's (2010, Experiment 2) shows a reproduction of the trends reported by Kiliç et al. (2017). These results were previously used to argue for an item noise account of the TPE where a differentiation process occurs through the course of testing such that strong items in the test list reduce item noise and mitigate the TPE. Our analyses instead demonstrated that context drift was the largest predictor of the TPE while item noise played very little role. These results further extent the viability and generalizability of a context drift account in addressing test list performance declines in recognition memory.

The fact that our model, which lacks a differentiation process, was capable of addressing the differential declines in pure strength and mixed strength lists shows that differentiation is not necessary to address such results. A comprehensive comparison between our model and the differentiation account of the TPE goes beyond the scope of this article, but one potential discriminating test might involve manipulations of the proportions of targets and lures in a test list. Targets are more likely to induce differentiation than lures, so a test list with more targets should result in more differentiation and thus a shallower decline in performance. A test list with more lures should result in less differentiation and thus a steeper decline in performance. Koop et al. (2015) compared the effects of testing when test lists contained all targets or all lures to test lists with half targets and half lures. When no feedback was present during testing, the declines in performance were virtually identical through the course of testing across the test list composition conditions. These results are seemingly contradictory to the differentiation account, but a model fit may be necessary to evaluate whether the model is capable of addressing the results.

12. The effect of task switching on recognition testing: data from Annis, Dube, and Malmberg (2016, Experiment 2)

A further puzzle for TPE models was reported by Annis et al. (2013), who demonstrated that items presented during a different task in the test phase had no influence on the TPE. Annis et al. (2013) compared two conditions. In one, participants had a long inter-stimulus interval (ISI) between trials, which we will refer to as the “blank” condition. In a second, “task-switching” condition, each test trial on the recognition test was followed by either a word or word-like stimulus presented for lexical decision (LD) if words were studied, or a face presented for gender identification (gender ID) if faces were studied. Surprisingly, both task-switching conditions produced nearly identical TPEs to the blank condition, despite the fact that the effective number of test stimuli is twice as large in the task-switching conditions.

Annis et al. (2013) argued that their data could be consistent with the results of a pure item noise model if items learned in the lexical decision trials were stored with a task-specific context representation that makes them resistant to interference during recognition testing. They additionally proposed that their results are contrary to contextual drift accounts, which they suggested would predict that LD trials should change the context representation used to cue memory, meaning that context drifts twice as quickly in the LD condition and thus should predict substantially more forgetting through the course of testing. Up to this point, the results of our modeling have suggested that context drift is principally responsible for declines in performance through recognition testing, but how could a context drift account explain the results of Annis et al. (2013)?

One consequence of task-switching manipulations that was not considered by Annis et al. (2013), however, is that there are often costs to the participants in terms of both performance and response time. Diffusion model analyses of task-switching conditions found that participants compensate by increasing their response boundaries, in addition to there being longer non-decision times and lower drift rates relative to conditions where participants continue to execute the same task (Karyanadis et al., 2009; Schmitz & Voss, 2012). Our model predicts that increases in response boundaries reduce the TPE. Using the same parameters as in Fig. 9, we generated model predictions for several different values of the response boundary parameter a with a sequence of 160 test trials – these predictions can be seen in Fig. 14. For the shallowest response boundary, the decline in HR was steep over the test list (Δ HR $= -.12$). For the most cautious response boundary, the decline in HR was considerably attenuated (Δ HR $= -.06$). This suggests that the task-switching manipulation may have incidentally introduced a factor that mitigated against the greater contextual change induced by the LD trials.

As the data from Annis et al. (2013) utilized the 2AFC paradigm, we applied the model to an analogous experiment utilizing the yes/no format conducted by Annis et al. (2016, Experiment 2). In fitting this dataset, we wanted to issue the strongest possible challenge to our model by implementing the suggested account by Annis et al. (2013). Specifically, items learned during the LD trials were assumed to be stored with a context specific to LD trials that does not match the context used to probe memory, and thus LD trials do not contribute to the total item noise. LD trials were assumed to generate context drift, so there was effectively twice as much context change in the LD condition as the blank condition.

In actuality, it may be possible that the uniqueness of the task context in LD might additionally induce less context change than

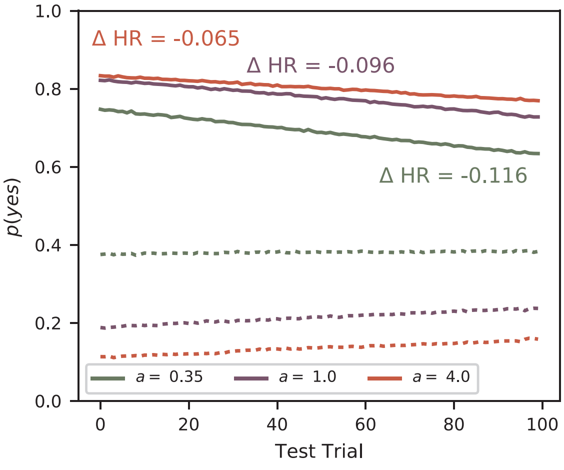

Fig. 14. Simulated predictions of HR and FAR from the model with three different intercepts for the response boundary parameter a.

recognition test trials. Nonetheless, at least one prominent model of episodic memory, namely the context maintenance and retrieval model (CMR: Polyn, Norman, & Kahana, 2009), assumes that task-switching causes a change in the episodic context in a similar manner to new items from within the same task. Similarly, investigations into the effect of testing on context change have found that semantic retrieval, which is occurring during the lexical decision trials, induces similar degrees of context change as episodic retrieval (Divis & Benjamin, 2014; Jang & Huber, 2008). For these reasons we fit the model with the assumption that the LD trials produce the same degree of context drift as the recognition test trials. One should additionally note that participants may have a considerable degree of difficulty maintaining a speed-accuracy threshold or bias as they alternate between the tasks, making this a plausible additional contender to explain the recognition memory results that was not considered by Annis et al. (2013).

One possibility is that the task-switching results of the previous investigations, which found higher speed-accuracy thresholds in task-switching conditions relative to controls, do not apply to the case of recognition memory. To demonstrate that the data of Annis et al. (2013) Experiment 2 requires higher speed-accuracy thresholds in task-switching conditions, we fit the standard DDM to their data and found evidence that participants employed higher speed-accuracy thresholds in the task-switching condition. These results can be found in the Supplementary Materials.

12.1. Parameterization

Task-switching manipulations produce effects on virtually all diffusion model parameters (Karyanadis et al., 2009; Schmitz & Voss, 2012). We followed this precedent and allowed t_{ER} and s_t, in addition to the linear functions on z/a and a, to vary across the blank and LD conditions. In addition, performance was considerably poorer in the LD condition relative to the blank condition. The most sensible assumption within the model to capture this decrement was to assume that participants might be cuing less effectively due to the rapid alternation the two tasks. We implemented this assumption by estimating the cue-to-target strength parameter, μ_{tt}, in the LD condition only. This parameter was also restricted to the (0, 1) interval for this condition, while it was fixed to 1 in the blank condition (and in all other fits).

We initially found this was not sufficient to capture the differences in the overall level of performance between the two conditions. Although performance is worse in the LD condition, there was also a much stronger bias to say “yes” and our initial fits found that the starting point z/a was not sufficient to capture this difference. For this reason, we additionally allowed the drift criterion, d_c, to vary across the LD and blank conditions, which yielded both a better fit and model selection score than the model with a single value of d_c across both conditions.[2]

A single learning rate parameter *r* was estimated for both conditions. Because the presentation rates were different from the Starns dataset, the learning rate parameter employed relatively uninformative priors. Similar to the previous datasets, a relatively short retention interval was used in this experiment and thus the context reinstatement parameter ξ was fixed to one. Because there was no word frequency manipulation in this dataset, ρ was again fixed to .039. Estimation of several model parameters was constrained by using the posterior distributions from the fit to Starns (2014, Experiment 2) as informed priors for this dataset. The complete list of parameters can be found in Table 3.

12.2. Posterior predictives

Table 4 indicates a model selection advantage for the $\sigma_{ti}^2 = 0$ model, which lacks item noise. However, here we decided to analyze the results of the full model, which includes the σ_{ti}^2 parameter, as the results from the task-switching paradigm were previously used to argue for an item noise account of test position effects. Fig. 15 depicts data and predictions for the experiment. Test lists were grouped into four blocks of 40 trials. The figure depicts choice probabilities (top) along with correct and error RTs (middle and bottom panels) averaged over participants. Due to the infrequency of errors ($M = 6.71$) relative to correct responses ($M = 12.47$) in each cell, each participant’s error RTs were summarized using the .25, .5, and .75 quantiles. Correct responses were summarized using the .1, .5, and .9 quantiles.

The data actually show *less* of a decrease in the HR in the LD condition (Δ HR $= -.14$ from the first to the last block) relative to the blank condition where there was no intervening task between recognition trials (Δ HR $= -.16$). This is surprising when one considers that when one includes the LD trials in the trial count, the effective number of test trials in the LD condition is twice as large (320) as the blank condition (160). Nonetheless, the model reproduces these trends, although it underpredicts the decline in performance, and notably predicts *less* of an HR decline in the LD condition than the blank condition ($M\Delta$ HR LD $= -.06$, $M\Delta$ HR blank $= -.12$). In contrast to previous datasets, the FAR decreases in the blank condition (Δ FAR $= -.04$), although the decline appears to be restricted to the final block. In the LD condition, the FAR does not show any consistent pattern. The model misses slightly here, predicting small increases in the FAR in each condition (Δ FAR blank $= .01$, Δ FAR LD $= .04$).

RTs were much longer in the LD condition than the blank condition, as evidenced by the shift in the RT distribution, and which the model captures. In terms of the effects of testing on RTs, this dataset differs from previous ones in that correct RTs appear relatively unaffected by the course of testing whereas error RTs decrease for targets in the blank condition and both targets and lures in the LD condition. The decrease in error RTs is consistent with the decrease in speed-accuracy threshold predictions in Fig. 3. We

[2] An additional possibility is that on some proportion of trials in the LD condition, participants ignored the task cue and responded as if they were performing lexical decision, saying “yes” to both targets and lures. However, full implementation of this idea would require a more complex mixture framework that models performance on the LD trials in addition.

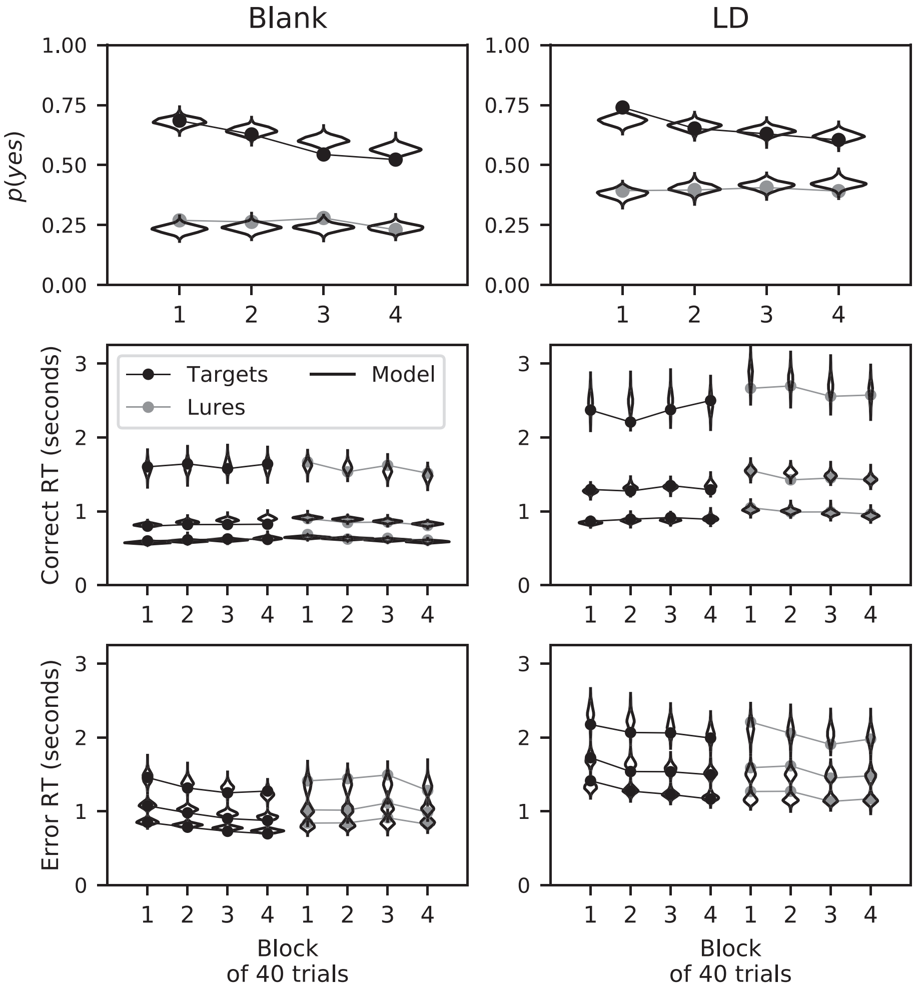

Fig. 15. Data and posterior predictive distributions for the data from Annis et al. (2016, Experiment 2) for each test block (40 trials per block) for the blank (left) and LD (right) conditions. Depicted are the choice probabilities (top) along with the correct (middle) and error (bottom) RTs. Correct RTs are summarized using the mean of each participant's .1, .5, and .9 quantiles of the RT distribution, while the error RTs are summarized using the .25, .5, and .75 quantiles. One should note that the test blocks in the LD condition refer to the recognition test trials.

demonstrate below that participants in this dataset deviated from previous datasets in that they demonstrated consistent decreases in their response boundaries through testing. Aside from the aforementioned deviations, the model appears to be providing a very good account of the data.

Plots and analyses of individual participants can be found in the Supplementary Materials, where it's shown that the model produces a good account of these data.

12.3. Analysis of parameter estimates

We have demonstrated that our model is capable of addressing the paradoxical pattern whereby lexical decision trials do not harm recognition memory performance, while additional recognition memory trials do. But do the conclusions from this dataset support the prior conclusions of the model, namely that context drift is the principle determinant of declines in performance through recognition testing?

Correlations between d'_{change} and the three major culprits of the TPE, namely the context drift parameter γ, the item mismatch variability parameter σ^2_{ti}, which governs the total degree of item noise, and the total change in response boundaries over the test sequence a_{change} (where $a_{change} = na_{slope}$, where n is the number of test trials), were calculated using the same procedure as before. Scatterplots along with 2D kernel density estimates of each of these comparisons for each experimental condition can be seen in Fig. 16.

Consistent with all previous analyses, the biggest predictor of the TPE was the context drift parameter γ, producing correlations of .75 and .64 in the blank and LD conditions, respectively. Again the item mismatch variability parameter, σ^2_{ti}, produced almost no correlation with the predicted TPE ($r \sim 0$). In contrast to prior analyses, changes in response boundaries produced only mild

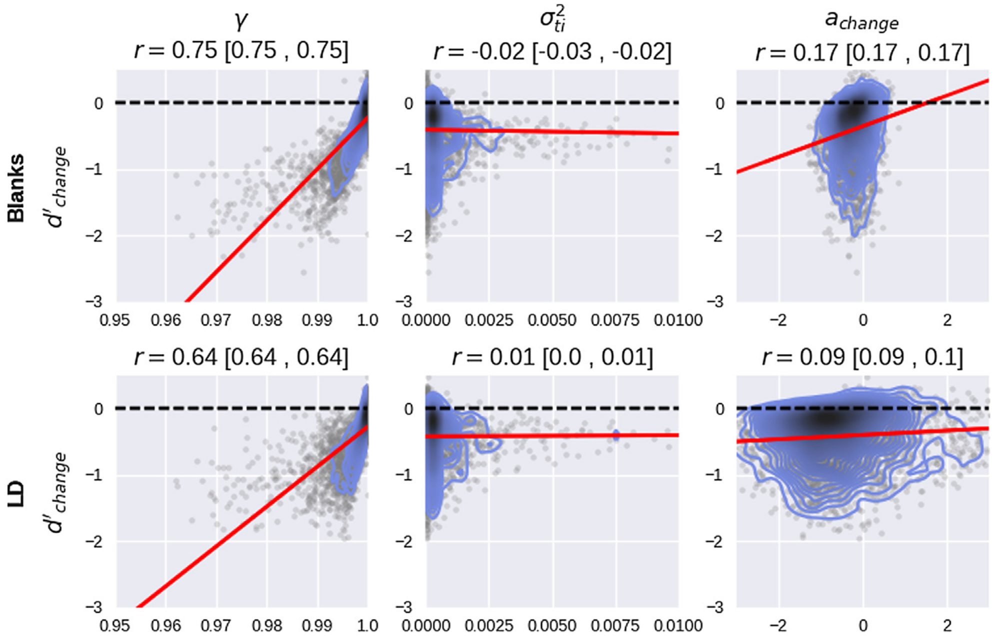

Fig. 16. Scatterplots (grey dots), 2D density estimates (shaded regions), and correlation coefficients (with upper and lower bounds of the 95% HDI reported in brackets) of the predicted d' decline through the test list against the parameters γ,σ^2_{ti}, and a_{change} for the Blank and LD conditions of Annis et al. (2016) Experiment 2. Red lines are linear regression estimates. Darker areas of the colored density depict the areas with the highest posterior density. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

correlations with the TPE, with correlations of .17 and .09 in the blank and LD conditions.

Histograms of the changes in bias and response boundary along with the intercept of the response boundary parameters can be seen in the fourth and fifth columns of Fig. 7. Participants in this dataset appeared to show a tendency to decrease both their bias and response boundaries through testing. 95% HDIs on the group parameters for the change in bias ($M_{blank} = -.14$, $[-.19, -.10]$, $M_{LD} = -.08$, $[-.12, -.05]$) and response boundaries ($M_{blank} = -.08$, $[-.12, -.05]$, $M_{LD} = -.30$, $[-.46, -.14]$) did not include zero. The very clear decreases in bias and speed-accuracy thresholds in this dataset are likely why the models that lack these components exhibited such large model selection penalties in Table 4. Participants also exhibited considerably higher intercepts on the response boundary function in the LD condition ($M = 2.32$, $[2.19, 2.45]$) than the blank condition ($M = 1.84$, $[1.74, 1.94]$), which is consistent with the analysis of the standard DDM in the Supplementary Materials.

12.4. Discussion

We implemented Annis et al. (2013)'s suggestion that LD trials cause context drift but do not increment item noise because the task context of the LD trials mismatches the recognition context. The model reproduced all the key trends in the data, including the critical trend that performance declined at roughly the same rate across the blank and LD conditions. Contrary to suggestions by Annis et al. that item noise principally drove the effect, we found that context drift was again the largest predictor of the TPE, despite the fact that the context drift process occurred at effectively twice the rate in the LD condition relative to the blank condition.

In this case, we found that decision dynamics were very influential, in that: (a) participants exhibited higher speed-accuracy thresholds in the LD condition than the blank condition to cope with the task-switching manipulation, and (b) participants exhibited clear decreases in both bias and speed-accuracy thresholds through the test. Contrary to previous datasets, error RTs noticeably decreased through testing, a pattern of data that requires changes in decision dynamics to address. This provides further support for the contention that RT data places important constraint on models of memory, as one might be mislead by choice proportions alone to conclude that the lack of change in performance is solely a memory phenomenon.

13. General discussion

A common finding across decades of recognition memory studies is that performance declines monotonically with increasing test position. In fact, the finding is so common that the three studies we chose to re-analyze in this work show robust effects of test position despite the fact that the studies were not designed to investigate the issue. Three hypotheses have been proposed to explain this phenomenon: (a) participants change their speed-accuracy thresholds through testing (b) the learned items exhibit interference and (c) the context used to cue memory changes through the course of testing.

Because the change in speed-accuracy thresholds is a decision level phenomenon, a memory model which is capable of addressing speed-accuracy tradeoffs is required to decisively measure these three influences. To this end, we extended the Osth and Dennis (2015) model by using the diffusion decision model (DDM) as a back-end decision model. The Osth and Dennis (2015) model is ideal for this purpose, as it is not only analytically tractable but can measure the contribution of item noise. Combined with the DDM, it is capable of addressing both the choices and RT distributions of benchmark recognition memory phenomena. We have also included a contextual change mechanism in the form of the Murdock (1997) contextual drift equation, which predicts an exponential decrease in performance across recognition test trials.

We have applied this new model to three major challenges for recognition memory models, namely: (a) the finding that the number of test items causes a greater decrement to performance than manipulations of the number of studied items, (b) the effects of study list composition on testing, with strong targets showing slower declines than weak items in pure lists (i.e.; all strong items or all weak items) but equivalent rates in mixed strength lists, and (c) the finding that intermixing lexical decision trials with recognition memory trials does not impair recognition memory performance. In order to constrain the relevant parameters governing memory retrieval, we constructed informed priors for each fit by using the posterior distributions of the group level parameters from the prior fit. We have additionally demonstrated that the model is capable of addressing a number of other benchmarks for recognition memory models, such as the effects of word frequency, study-test delay, and list length on both choice probabilities and response times.

A consistent trend emerged from the analysis of the parameters from each dataset: changes in the retrieval context were most predictive of the overall decline in recognition memory performance, whereas item noise played only a small role in the decline. Changes in speed-accuracy thresholds additionally showed strong correlations with the change in response boundaries. However, in most of the analyses the mean of the change was around zero, but there was a high degree of variability across participants, with some increasing their speed-accuracy thresholds through the test and others decreasing them. One exception was the fit to the data from Annis et al.'s (2016) experiment, where there were consistent decreases in response boundaries in both conditions, but in this dataset the decrease exhibited a relatively small correlation with the TPE. Thus, it appears that changes in response boundaries play a role in moderating the size of the TPE due to the large individual differences across participants, but this factor does not appear to be sufficient to drive the accuracy effect on its own. The results of a parameter removal analysis presented in the Supplementary Materials, which demonstrates the predicted effects of each culprit individually, support the conclusions above and demonstrated that changes in response boundaries through the test produced almost no decrement through testing on average. In addition, these trends did not just emerge from the full model, but were also present across a number of alternative models with alternative parameterizations (these results are presented in the Supplementary Materials).

Criss et al. (2011) criticized context change as an ad hoc explanation of the TPE. We argue that the situation is exactly the opposite. Many independent reports have suggested that retrieval is one of the largest sources of context change (Divis & Benjamin, 2014; Jang & Huber, 2008; Klein et al., 2007; Pastötter et al., 2011; Sahakyan & Hendricks, 2012; Sahakyan & Smith, 2014). Klein et al. (2007) presented compelling evidence from a two-alternative forced choice (2AFC) recency discrimination task. Participants had to choose the more recent of the two items, and activity between the two items was varied. Performance was close to chance when the activity was studying new faces or performing math problems, but was much higher when a recognition test intervened, suggesting that context changed to a greater degree during the recognition test. Converging evidence also comes from other memory models. A number of free recall models, such as the model of Davelaar, Goshen-Gottstein, Ashkenazi, Haarmann, and Usher (2005) and variants of the temporal context model (Howard & Kahana, 2002; Polyn et al., 2009; Sederberg et al., 2008), posit that each retrieval event changes the context used to cue memory. In addition, Jonker, Seli, and MacLeod (2013) were able to account for the major findings in the literature on retrieval-induced forgetting with assumptions about retrieval during the retrieval practice phase causing context change. Thus, there appears to be strong independent evidence for the assumption that retrieval causes context change.

The mechanism of context drift has a lot of explanatory power outside of its effects on recognition testing. Context drift allows for the explanation of recency effects at both short and long time scales (Howard & Kahana, 1999; Sederberg et al., 2008), lag effects in continuous recognition (Murdock, 1997), spontaneous recovery (Estes, 1955), spacing effects (Glenberg, 1976; Siegel & Kahana, 2014), and contiguity effects (Howard & Kahana, 1999; Polyn et al., 2009; Sederberg et al., 2008), along with explanations of how prior studied or tested lists impact performance on the current list (Lohnas, Polyn, & Kahana, 2015; Mensink & Raaijmakers, 1988). Given the wealth of evidence supporting the mechanism of context drift, it appears quite plausible that this mechanism plays a role in the TPE.

While we did not find evidence that item noise causes decrements in recognition memory testing, it is important to note that our results did not indicate that there is no item noise in recognition memory, contrary to previous suggestions (e.g.: Dennis & Humphreys, 2001). For each dataset, we conducted a model selection exercise where we implemented simpler models that lacked one of the major culprits of the TPE. Elimination of item noise caused an impairment in model selection in two of the four datasets. In addition, the analysis of our new experiment (Experiment 1) showed that item noise played a larger role in explaining the list length effect than it did in explaining the TPE. One reason why item noise was not able to explain both phenomena is because the model predicts that the TPE should be smaller than the list length effect, as shown in Fig. 3, which stands in contrast to the data. This is because of the inverse square root relationship between the number of items in memory and d'; as more items are added to memory, each item exhibits less and less of a detriment to total performance. Thus, the increase in list length between 24 and 96 items should exhibit more of a decline in performance than between 24 and 96 test trials, where the number of items in memory increases from 120 at the 24th test trial to 192 by the 96th test trial. However, it should also be mentioned that models such as REM may make different predictions than our model, especially given that differentiation operating during the test phase can produce unique predictions (e.g.; Kiliç et al., 2017) and thus may allow for the prediction of a dissociation between list length and test position effects.

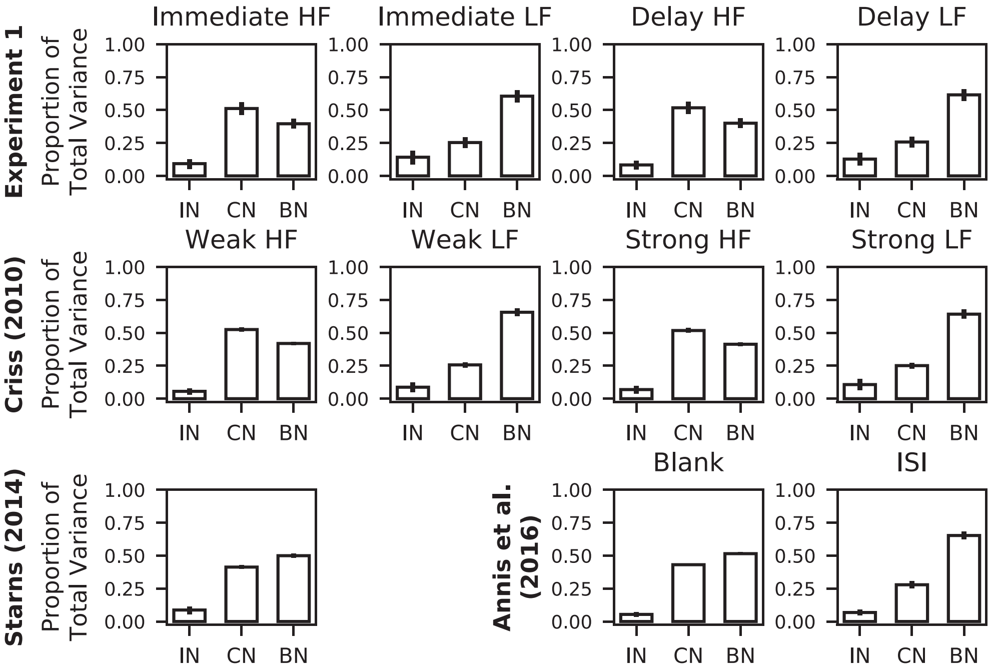

Fig. 17. Barplots depicting the proportion of total variance occupied by item noise (IN), context noise (CN), and background noise (BN) for Experiment 1 (top row), the data from Criss (2010, Experiment 2, middle row), the data from Starns (2014, Experiment 2, bottom left corner), and the data from Annis et al. (2016, Experiment 2, bottom right). Error bars depict the 95% highest density interval (HDI).

In our previous work, we argued that item noise plays only a small role in recognition memory performance, which was based on the finding that the total item noise contribution was substantially less than that of pre-experimental sources (context noise in the case of HF words and background noise; Osth & Dennis, 2015). We reproduced this analysis for the present work by computing the item noise and context noise contributions according to Eq. (10) for each dataset using the group mean parameters, with results depicted in Fig. 17. For Experiment 1, only the long list results are shown, and each case depicts the interference after the final test trial in the test sequence. Because background noise and the mean context noise were fixed in many of the analyses, the results were depicted as the proportions of total variance for memory strength. The trends in Fig. 17 largely reproduce those found by Osth and Dennis (2015), in that background noise and context noise are largely responsible for HF words whereas LF words are more dominated by background noise (due to their lesser amount of context noise).

Our previous work found that item noise was almost negligible for word stimuli, whereas here we found that mean item noise proportion became as large as 14.2% for LF words in the immediate condition of Experiment 1. This difference can be explained because the present modelling differs in important ways. First, we model response times, which has been shown to introduce considerable constraint on cognitive models. Second, we modeled changes in performance through the course of testing. Finally, Osth and Dennis (2015) also included results from the associative recognition task, where the effects of strength on the FAR to rearranged pairs places further constraint on the parameter as item noise predicts larger FAR for rearranged pairs as strength is increased (Osth & Dennis, 2014). These additional constraints may have decreased the magnitude of the estimated contribution of item noise. Nonetheless, our results are qualitatively consistent with the idea that item noise does not underpin the bulk of interference in recognition memory. Osth and Dennis (2015) argued that these results can be understood if item representations are largely dissimilar to each other, which is consistent with both empirical and computational work showing that the hippocampus exhibits sparse representations (e.g.: Marr, 1971; McClelland, McNaughton, & O'Reilly, 1995; Norman & O'Reilly, 2003).

13.1. Changes in bias and speed-accuracy thresholds

One surprising result of our analyses was the consistently large individual variation in participants' changes in response bias and speed-accuracy thresholds. In addition, reduced models that lacked such changes in decision dynamics suffered very large WAIC penalties relative to removal of the components related to memory retrieval. Although the response boundary parameters did not show consistent trends upward or downward, response bias parameters showed consistent decreases in many of the comparisons, which replicates the findings of Ratcliff (1978), who found decreases in starting point over blocks of recognition testing.

What might be causing such changes? Insight here could be achieved by understanding some of the mechanism behind trial-by-trial adjustments of such parameters. Analyses with the DDM have found that participants often increase their speed-accuracy threshold and adjust their bias in response to feedback that they had received an error, which are associated with slowing on subsequent trials (post-error slowing; Dutilh et al., 2012). No such feedback was present in our experiments, and thus participants were unlikely to know whether they made an error. One model that predicts trial-by-trial changes in such parameters is the self-regulating accumulator model (Lee & Dry, 2006; Vickers, 1979; Vickers & Lee, 1998), which determines choice, confidence, and RT from a race between two accumulators. Although choice and RT are determined by the winning accumulator, confidence is determined by the difference in evidence between the winning and losing accumulator. In the self regulating accumulator model, modulations in the response thresholds for each accumulator occur on a trial-by-trial basis in order to maintain a target level of confidence. Thus, one possibility is that the changes in performance through testing, which produce bigger decreases in performance for target items, reduce confidence to a larger extent for targets, and participants aim to compensate for this reduction by increasing the threshold of the accumulator corresponding to the "old" response, which in turn produces a bias towards new responses.

An interesting possibility would be to extend the current model to confidence ratings and evaluate whether the model could generate the required changes in bias and boundaries with self regulating accumulators. One of the difficulties in applying self regulation is its reliance on the balance-of-evidence mechanism to determine confidence. The currently dominant models of RT, choice, and confidence are the response time confidence models (RTCON and RTCON2: Ratcliff & Starns, 2009, 2013), which are accumulator models where each confidence option receives its own accumulator. Ratcliff and Starns (2009) argued an advantage of this approach is the finding that RT distributions for different levels of confidence can be quite similar under some circumstances, a result that seems superficially at odds with a balance-of-evidence mechanism, which would predict faster high confidence responses. Although it might be possible to replace the DDM with RTCON2 as the back-end decision process in our model, it remains to be seen whether the dynamics of self regulation can be built into the RTCON models, as these models are already quite complex and use many more parameters than the relatively simple balance-of-evidence models.

13.2. Ratios of target to lure variability

A near universal finding in the ROC literature is that targets have higher variability than lures (Heathcote, 2003; Ratcliff et al., 1992; Wixted, 2007). Osth and Dennis (2015) demonstrated that their model was capable of addressing ROC shapes when the item match variability parameter σ_{tt}^2 was higher than the item mismatch variability parameter σ_{ti}^2. This means that the match between an item cue and its own representation has higher variability than the match between a cue and another item in memory, which is possible if there is both encoding variability (producing match variability) and sparse representations (producing low mismatch variability). This can be seen in the expression for target variance (Eq. (9)), which differs from the equation for lure variance (Eq. (10), both equations are present in Appendix A) because of the self match term in Eq. (9) that contains σ_{tt}^2. Although we lack any datasets with a bias manipulation which would enable us to construct an ROC curve, we are able to constrain the estimate of target-to-lure variability using response times, as has been demonstrated with numerous diffusion model fits that have found higher drift rate variability for targets than for lures (Osth, Dennis, et al., 2017; Osth, Bora, et al., 2017; Starns & Ratcliff, 2014; Starns et al., 2012; Starns, 2014).

We constructed estimates of the target-to-lure variability S (i.e., the inverse of zROC slope) using Eqs. (9) and (10) with our estimated group mean parameters in each dataset. The results are depicted in Fig. 18. Because S changes over the course of testing, we depict the estimates for the first trial only. As expected the estimates of S are higher than 1 in all conditions, consistent with zROC slopes less than one. Additionally, similar to recognition memory models such as REM (Shiffrin & Steyvers, 1997) and BCDMEM (Dennis & Humphreys, 2001), our model predicts higher estimates of S for conditions of higher performance (LF words, strong items, etc.) despite using a single estimate of σ_{tt}^2 in each of these conditions. An exception to the rule is the LD condition of the Annis et al. (2016) dataset, where performance is worse by virtue of a lower value of the cue-to-target strength parameter μ_{tt}, which increases the S estimates.

Eqs. (9) and (10) reveal how the model produces higher variability in conditions of higher performance in some cases but not others. Consider if one were to manipulate the context mismatch variability parameter ρ to capture different levels of word

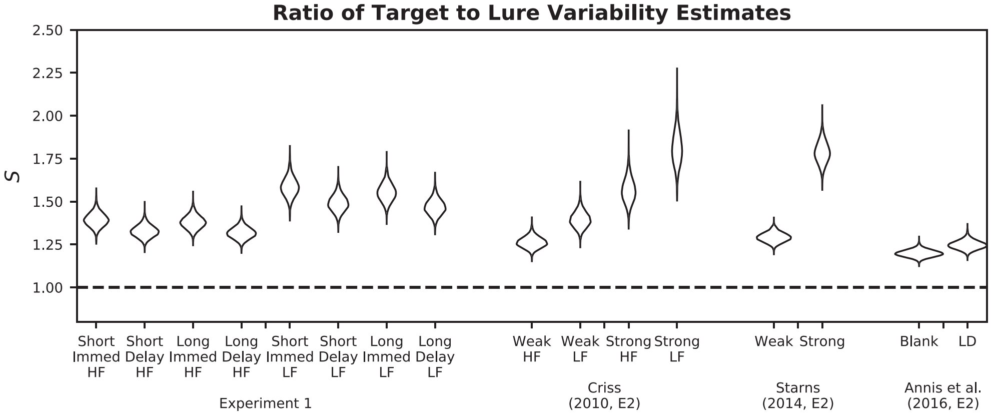

Fig. 18. Violin plots showing the posterior distribution of target-to-lure variability S for each condition constructed from the group mean parameters.

frequency. This has the effect of increasing the context noise variability, which increases the variability for targets and lures by the same amount, decreasing the ratio. Decreases in μ_{tt}, in contrast, decrease the context noise variance (affecting both targets and lures) but do not affect the self match term when $\sigma_{ss}^2 = 0$ (which was the case in our fits). Because the variability of both distributions decreases by the same amount when μ_{tt} decreases, this results in an increase in the ratio of target-to-lure variability S seen in the Annis et al. (2013) data.

These results stand in contrast to analyses with the DDM, which have been unable to detect differences in S across different performance manipulations, such as word frequency, strength, or speed-accuracy emphasis (Osth, Bora, et al., 2017; Starns, 2014; Starns & Ratcliff, 2014). However, inspection of Fig. 18 reveals that the uncertainty in the estimates of S is quite large, which has also been demonstrated with the standard DDM (Osth, Bora, et al., 2017; Osth, Dennis, et al., 2017). Second, although models that freely estimate S do not find differences in S across conditions, models such as ours, which impose a relationship between S and the mean drift rate, still appear to be able to provide a good account of RT distributions across targets and lures. In addition, analyses of confidence and RT together have found evidence that LF words have higher variability than HF words (Ratcliff & Starns, 2009).

13.3. Comparison to feature sampling models of recognition memory

Our model resembles the exemplar-based random walk model (EBRW: Nosofsky et al., 2011) and exemplar-based linear ballistic accumulator (EBLBA: Donkin & Nosofsky, 2012a) models of short-term recognition memory, in the sense that there is a memory retrieval front-end that generates evidence for a back-end decision model that generates predictions of choices and response times. An alternative framework for predicting response times is the class of feature sampling models (Brockdorff & Lamberts, 2000; Cox & Shiffrin, 2012; Malmberg, 2008). In feature sampling models, RT is determined by a dynamic process where the cue vector begins with only a limited number of features. On each iteration of retrieval, features are sampled randomly with replacement and memory strength is calculated via a global matching process. Retrieval stops when either the memory strength falls above a "yes" criterion or below a "no" criterion.[3] The number of iterations determines the response time.

An advantage of feature sampling models is their ability to predict changes in performance as retrieval unfolds, and which cannot be addressed with a simple change in the speed-accuracy threshold. A compelling demonstration of this was given by Hintzman and Curran (1994), who compared old items, unrelated lures, and switched plurality lures (e.g.: if "cat" was studied, "cats" would be a switched plurality lure). Hintzman and Curran used the signal-to-respond procedure, where retrieval was interrupted by signals at various time lags following stimulus presentation to demand a response from the participant, resulting in an acceptance rate as a function of time for each class of items. Although HR rose monotonically and FAR to unrelated lures showed monotonic decreases, switched-plurality lures showed a non-monotonic FAR function, where FARs rose initially but fell at later retrieval times (a finding replicated by Rotello & Heit, 1999). Brockdorff and Lamberts (2000) were able to model this result by assuming that the features that represent whether a stimulus is singular or plural are sampled later than the other stimulus features, making it such that targets and switched-plurality lures are more similar early in retrieval but become distinguished later in retrieval when the plurality features are sampled.

Aside from this finding, however, there are few other findings in the literature on single item recognition memory that require dynamic, non-stationary evidence. Several of the manipulations investigated in this article have been similarly investigated using the signal-to-respond paradigm. When d' is calculated as a function of retrieval lag, the speed-accuracy tradeoff (SAT) function can be characterized using an exponential rise to asymptote (e.g.: Reed, 1973). This function has three parameters: an on-set parameter that describes when the function begins to rise, asymptotic d', and the rate of rise to the asymptote. Analyses with the DDM have found that changes in drift rate affect only the asymptotic d' of the SAT function, whereas changes in the rate parameter of the SAT function require a change to either the diffusion noise or require a non-stationary process, such as the aforementioned feature sampling models (Ratcliff, 1978, 2006). The variables investigated in this article such as word frequency (Hintzman, Caulton, & Curran, 1994), strength (Dosher, 1984; Kiliç & Öztekin, 2014), study-test delay for lags longer than immediate repetition (Dosher, 1981; McElree & Dosher, 1989), and list length (McElree & Dosher, 1989; Reed, 1976), affect only the asymptote of the SAT function, suggesting that they can be modeled by a stationary process. Similarly, Gillund and Shiffrin (1984) manipulated list length, item strength, and whether participants were forced to respond slowly or quickly. Better performance was found for slow responses in all conditions, but there were no interactions with list length or presentation time, suggesting there was no unique contribution late in retrieval. Thus, a model suchy as ours which contains stationary evidence through the course of the decision may be appropriate for many cases in single item recognition.

14. Conclusions

We have constructed a new integrated model of choice and response time in recognition memory which inherits the retrieval from memory component of the Osth and Dennis (2015) model and the diffusion decision model to address how memory strength produces decisions and the time it takes to make them. Such a combined architecture has extra utility when one considers the criticisms of evidence accumulation models by Jones and Dzhafarov (2014), who argued that drift rate distributions have an arbitrary form and the models are unfalsifiable if the distributional form is allowed to vary over conditions. In our model, in contrast, the distributions of

[3] The Cox and Shiffrin (2012) model deviates from this somewhat. In this model, on each iteration the difference in memory strength between the current iteration and the previous iteration is calculated, with positive increments driving an "OLD" counter and negative increments driving a "NEW" counter.

drift rates were generated by the memory model front-end rather than being estimated as free parameters. We applied the combined model to several datasets that show impairments in recognition memory as a function of the number of test trials, revealing that both memory strength and decisional factors are responsible. This work demonstrates the utility, and indeed often the necessity, of considering the dynamics of decision making in recognition memory. Our work also suggest that contextual drift mechanisms are under-explored in recognition memory. Indeed, many of the phenomena that have been argued to support such mechanisms, such as spacing effects (Glenberg, 1976) and long-term recency effects (Talmi & Goshen-Gottstein, 2006) have been observed in recognition memory in addition to recall tasks.

Author note

Model code and supplementary materials can be found at https://osf.io/5kghv/. We would like to thank Amy Criss, Jeff Starns, and Jeff Annis for generously providing their data and Daniel Little for helpful comments and discussions. This work was supported by a grant from the Australian Research Council, ARC DP150100272, awarded to Simon Dennis and Andrew Heathcote and ARC DE170100106 awarded to Adam Osth.

Appendix A. Analytic derivation of the Osth and Dennis (2015) model

Following Humphreys, Pike, Bain, and Tehan (1989), we can deconstruct Eq. (2) into the various components that comprise the memory matrix M, assuming that memory is being probed for a target item:

$$\begin{aligned} s = (C_s' \otimes I_t')\cdot[r(C_s \otimes I_t) & \qquad \text{Self Match} \\ + \sum_{i\in L, i\neg=t} r(C_s \otimes I_i) & \qquad \text{Item Noise} \\ + \sum_{u\in P, u\neg=s} (C_u \otimes I_t) & \qquad \text{Context Noise} \\ + \sum_{u\in P, u\neg=s, z\notin L} (C_u \otimes I_z)] & \qquad \text{Background Noise} \end{aligned} \tag{3}$$

where i indicates items on the study list (L), that are not the item cue t, the u subscript indicates a prior list context from the set of all contexts prior to the study list (P), and z indicates items from prior list contexts that were not on the study list.

Eq. (3) can then be rewritten as the match between the cue vectors and the stored vectors:

$$\begin{aligned} s = r(C_s'\cdot C_s)(I_t'\cdot I_t) + & \qquad \text{Self Match} \\ \sum_{i\in L, i\neg=t} r(C_s'\cdot C_s)(I_t'\cdot I_i) + & \qquad \text{Item Noise} \\ \sum_{u\in P, u\neg=s} (C_s'\cdot C_u)(I_t'\cdot I_t) + & \qquad \text{Context Noise} \\ \sum_{u\in P, u\neg=s, z\notin L} (C_s'\cdot C_u)(I_t'\cdot I_i) & \qquad \text{Background Noise} \end{aligned} \tag{4}$$

In this form the three sources of interference (item noise, context noise, and background noise) are described as matches and mismatches between the item and context vectors. These dot products can be parameterized using normal distributions:

$$\begin{aligned} C_s'\cdot C_s &\sim Normal(\xi,0) & \text{Context Match} \\ C_s'\cdot C_u &\sim Normal(0,\sigma_{su}^2) & \text{Context Mismatch} \\ I_t'\cdot I_t &\sim Normal(\mu_{tt},\sigma_{tt}^2) & \text{Item Match} \\ I_t'\cdot I_i &\sim Normal(0,\sigma_{ti}^2) & \text{Item Mismatch} \end{aligned} \tag{5}$$

The means and variances of the distributions of dot products are the parameters of the model. This approach is similar to the kernel trick employed by support vector machines (Schölkopf & Smola, 2002). The choice of the normal distribution offers mathematical convenience for this application by allowing separate specification of the mean and variance parameters. Covariances were avoided by fixing the means of the mismatch distributions to zero. Fixing the variability of the context match to zero avoids covariances during the context drift process. In our prior work, we allowed for variability in the context match and it exhibited a similar effect as the item match variability, in that it allowed for greater variability for targets than lures (Osth & Dennis, 2015). We demonstrate in the main text that the model is similarly able to produce reasonable estimates of target variability even with no context match variability.

The distributions of the matches and mismatches from Eq. (5) are substituted into the terms for Eq. (4) to derive mean and variance expressions for the signal and noise distributions. Because each noise term is the multiplication of an item match/mismatch by a context match/mismatch, and each is represented by a normal distribution, each term is a multiplication of normal distributions, which results in a modified Bessel function of the third kind with mean and variance as follows:

$$E(X_1X_2) = \mu_1\mu_2$$
$$V(X_1X_2) = {\mu_1}^2\sigma_2^2 + {\mu_2}^2\sigma_1^2 + \sigma_1^2\sigma_2^2$$

Given the large number of list items and non-list items that are stored in the occurrence matrix, the final distribution of memory strength is the sum of many product distributions and the sum is approximately normal by virtue of the central limit theorem. The mean and variance for the old (target) and new (lure) distributions are as follows:

$$\mu_{old} = r\xi\mu_{tt} \tag{6}$$

$$\mu_{new} = 0$$

$$\begin{aligned}\sigma_{old}^2 = r^2(\xi^2\sigma_{tt}^2)+ &\quad \text{Self Match}\\ r^2(l{-}1)(\xi^2\sigma_{ti}^2)+ &\quad \text{Item Noise}\\ m(\mu_{tt}^2\sigma_{su}^2 + \sigma_{su}^2\sigma_{tt}^2)+ &\quad \text{Context Noise}\\ n(\sigma_{su}^2\sigma_{ti}^2) &\quad \text{Background Noise}\end{aligned} \tag{7}$$

$$\begin{aligned}\sigma_{new}^2 = r^2l(\xi^2\sigma_{ti}^2)+ &\quad \text{Item Noise}\\ m(\mu_{tt}^2\sigma_{su}^2 + \sigma_{su}^2\sigma_{tt}^2)+ &\quad \text{Context Noise}\\ n(\sigma_{su}^2\sigma_{ti}^2) &\quad \text{Background Noise}\end{aligned} \tag{8}$$

where l is the length of the list, m is the number of pre-experimental memories of the target item, and n is the total number of background memories. The rows of Eq. (7) can be viewed as the contributions of the self match, item noise, context noise, and background noise. Eqs. (7) and (8) are identical with the exception of the self match variance term, which is only in Eq. (7), and the fact that item noise is scaled by $l{-}1$ in Eq. (8) instead of l.

We can further simplify the model by combining the m and σ_{su}^2 terms into a the parameter ρ which reflects the total context mismatch variability. Additionally, we eliminate the entire background noise term and instead substitute a separate variance parameter to reflect its contribution, which we denote as β. Finally, we will substitute all mean context match terms ξ with a new term ω that reflects the context similarity after context drift has occurred (see the section below for a derivation of ω). The simplified variance equations are as follows:

$$\begin{aligned}\sigma_{old}^2 = r^2(\omega^2\sigma_{tt}^2)+ &\quad \text{Self Match}\\ r^2(l{-}1)(\omega^2\sigma_{ti}^2)+ &\quad \text{Item Noise}\\ (\mu_{tt}^2\rho + \rho\sigma_{tt}^2)+ &\quad \text{Context Noise}\\ \beta &\quad \text{Background Noise}\end{aligned} \tag{9}$$

$$\begin{aligned}\sigma_{new}^2 = r^2l(\omega^2\sigma_{ti}^2)+ &\quad \text{Item Noise}\\ (\mu_{tt}^2\rho + \rho\sigma_{tt}^2)+ &\quad \text{Context Noise}\\ \beta &\quad \text{Background Noise}\end{aligned} \tag{10}$$

Although list length manipulations can be implemented by changing the value of l in Eqs. (9) and (10), learning through the course of testing is complicated by the fact that context drift is occurring, making it such that recently acquired test items have a stronger match to the context cue than older test items or study list items. For this reason, items learned through the course of testing are modeled by including separate item noise terms (second row in Eqs. (9), first row in Eq. (10)) for each tested item, where $l = 1$. When each tested item is initially added to memory, its context match ω is 1 and subsequently decays as a consequence of context drift (see The Context Drift Equation subsection for more details).

To improve sampling, an average value ρ_{avg} along with an effect size parameter ρ_{wf} were sampled. ρ_{wf} is defined as a proportion of ρ_{avg} that differs between and HF and LF words and is bounded on a $[0, 1]$ interval. ρ_{LF} and ρ_{HF} were defined as follows:

$$\rho_{LF} = \rho_{avg}-(\rho_{avg}\rho_{effect}) \tag{11}$$

$$\rho_{HF} = \rho_{avg} + (\rho_{avg}\rho_{effect}) \tag{12}$$

A.1. The likelihood ratio transformation

The log-likelihood ratio λ for a given class of items can be calculated usingthe linear approximation to the log-likelihood transformation devised by Osth, Dennis, et al. (2017), which results in normal distributions of log likelihood ratios. These expressions were written for the general case in terms of discrimination d and the relative variability of the target distribution S, which we can reach by normalizing the parameters by σ_{new}:

$$d = \mu_{old}/\sigma_{new} \tag{13}$$

$$S = \sigma_{old}/\sigma_{new} \tag{14}$$

The means and standard deviations of λ can be expressed in terms of d and S resulting in normal distributions with the following means and standard deviations:

$$\mu_{\lambda new} = -((\frac{d^2}{2})(\frac{S^2+3}{4S^2}) + log(S)) \tag{15}$$

$$\mu_{\lambda old} = d^2\frac{S^2+1}{2S^2} + \mu_{\lambda new} \tag{16}$$

$$\sigma_{\lambda new} = d\frac{S^2+1}{2S^2} \tag{17}$$

$$\sigma_{\lambda old} = S\sigma_{\lambda L} \tag{18}$$

In conditions where participants study a mixed strength study list, such as when items are studied either once or three times, then if they are not given a cue as to which strength they will be tested on, it is unreasonable to assume that the expected strengths used in the likelihood transformation correspond to the true strengths. For instance, if a participant studied such a mixed list and was tested on an item that was studied three times, the participant would have to first know that the item was a studied in order to know that it was studied three times. For this reason, it is more appropriate for the expected strengths in the likelihood ratio transformation of a mixed list to be the average of the strong and weak item strengths (Osth & Dennis, 2015; Starns et al., 2010). This can be accomplished by averaging the learning rates from the two strength conditions to generate r_{avg} and then generating the expected strengths d and S according to the above equations. The actual learning rates r_{weak} and r_{strong} are used to generate the actual strengths for a given condition, which we denote as d^* and S^*. Expressions for the target distributions of a mixed strength case are thus:

$$\mu_{\lambda old} = dd^*\frac{S^2+1}{2S^2} + \mu_{\lambda L} \tag{19}$$

$$\sigma_{\lambda old} = S^*\sigma_{\lambda L} \tag{20}$$

The lure expressions for a mixed list are unchanged.

A.2. The context drift equation

Following Murdock (1997), we begin by assuming that a context vector C on trial j is a weighted combination of a new context z and the context on the previous trial i:

$$C_j = \gamma C_i + (\sqrt{1-\gamma^2})z \tag{21}$$

where γ is a scalar ranging between zero and one that provides the weighting on the context from the previous trial. If $\gamma = 1$, context does not change from trial to trial. If we further adopt the simplifying assumptions that $E[C_i.\ z] = 0$ and $Var[C_i.\ z] = 0$,[4] the new context z no longer plays a role and we can express the overall context match ω on trial j as:

$$\omega_j = \gamma\xi \tag{22}$$

Since the variability of the context match is fixed to zero, the amount of contextual drift that has occurred only affects the mean of the context match. This can be expressed more generally for a given lag l after trial i as:

$$\omega_l = \gamma^l\xi \tag{23}$$

where ξ behaves as an intercept that decays exponentially over successive trials. For all immediate testing conditions, ξ is fixed to 1 for studied items for the first test trial and decays over successive trials. For studied items in delayed testing conditions we fit ξ as a free parameter to capture the match to context on the first trial, a value which varies between 0 and 1 to reflect the idea that the match to context is always weaker in delayed testing relative to immediate testing.

Appendix B. Prior distributions on model parameters and details of the hierarchical modeling procedure

Participant parameters are sampled from group level mean and standard deviation parameters M and ς. Following previous examples with the DDM, bounded parameters were sampled from truncated normal distributions:

[4] This assumption is met if the two context vectors were orthogonal to each other (e.g., Howard & Kahana, 2002).

$$
\begin{aligned}
z/a_{int.} &\sim TN(M_{zint}, \varsigma_{zint}, 0, 1)\\
z/a_{slope} &\sim Normal(M_{zslope}, \varsigma_{zslope})\\
a_{int.} &\sim TN(M_{aint}, \varsigma_{aint}, 0, \infty)\\
a_{slope} &\sim Normal(M_{aslope}, \varsigma_{aslope})\\
t_{er} &\sim TN(M_{ter}, \varsigma_{ter}, 0, \infty)\\
s_t &\sim TN(M_{st}, \varsigma_{st}, 0, \infty)\\
d_c &\sim Normal(M_{dc}, \varsigma_{dc})
\end{aligned}
$$

Because the z/a is a proportion and its constituents are positive, z/a falls between zero and one and was truncated to the (0,1) interval. Other parameters, such as a,t_{er}, and s_t are bounded below at 0 but unbounded on the right. The slope parameters z/a_{slope} and a_{slope} can be positive or negative, and were thus sampled from normal distributions.

We followed the parameterization of Osth and Dennis (2015) for their model, where parameters that were lower bounded at zero but unbounded on the right were sampled from normal distributions on a log scale:

$$
\begin{aligned}
\log(\sigma_{ti}^2) &\sim Normal(M_{\sigma ti}, \varsigma_{\sigma ti})\\
\log(\sigma_{tt}^2) &\sim Normal(M_{\sigma tt}, \varsigma_{\sigma tt})\\
\log(\rho) &\sim Normal(M_{\rho}, \varsigma_{\rho})\\
\log(r) &\sim Normal(M_r, \varsigma_r)
\end{aligned}
$$

An exception is the ρ_{effect} parameter, where $\rho_{HF} = \rho + (\rho_{effect}\rho)$ and $\rho_{LF} = \rho - (\rho_{effect}\rho)$. Because ρ_{effect} defines the effect size of the word frequency effect as a proportion of ρ, it is bounded on the (0, 1) interval and is sampled as follows:

$$\rho_{effect} \sim TN(M_{\rho effect}, \varsigma_{\rho effect}, 0, 1)$$

Following Osth and Dennis (2015), some of the parameters on the (0, 1) interval were sampled from beta distributions. Improved sampling for the beta distribution can be obtained by reparameterizing it in terms of its mean μ and sample size v:

$$
\begin{aligned}
\alpha &= \mu v\\
\beta &= (1-\mu)v
\end{aligned}
$$

We will henceforth denote the reparameterized beta distribution as *rBeta*.

The following parameters were sampled from reparameterized beta distributions:

$$
\begin{aligned}
\gamma &\sim rBeta(\mu_{\gamma}, v_{\gamma})\\
\xi &\sim rBeta(\mu_{\mu ss}, v_{\mu ss})\\
\mu_{tt} &\sim rBeta(\mu_{\mu tt}, v_{\mu tt})
\end{aligned}
$$

For the group level mean (*M*) parameters of the DDM, we used mildly informative priors, several of which were employed by Osth, Dennis, et al. (2017):

$$
\begin{aligned}
M_{ter} &\sim TN(.5, .5, 0, \infty)\\
M_{st} &\sim TN(.25, .25, 0, \infty)\\
M_{z/aint} &\sim TN(.5, .5, 0, 1)\\
M_{aint} &\sim TN(2, 2, 0, \infty)\\
M_{z/aslope,aslope} &\sim Normal(0, .05)\\
M_{dc} &\sim Normal(0, 1)
\end{aligned}
$$

For the *M* parameters of the Osth and Dennis (2015) model, we employed their non-informative priors for Experiment 1:

$$
\begin{aligned}
M_{\sigma ti,\sigma tt,\rho,r} &\sim Normal(0, 100)\\
M_{\gamma,\mu ss,\mu tt} &\sim rBeta(.5, 2)
\end{aligned}
$$

One should note that *rBeta* with $\mu = .5$ and $v = 2$ is equivalent to a beta distribution with $a,b = 1$, where all values on the (0, 1) interval have equal probability mass. For the $M_{\rho effect}$ we used:

$$M_{\rho effect} \sim TN(.5, .5, 0, 1)$$

For the group level standard deviation (ς) parameters of the DDM we used the following mildly informative priors:

$$
\begin{aligned}
\varsigma_{aint,dc} &\sim Gamma(1, 1)\\
\varsigma_{z,st,t0,aslope,zslope} &\sim Gamma(1, 3)
\end{aligned}
$$

For the ς parameters of the Osth and Dennis (2015) model, we used less informative priors on several of the parameters:

$$\begin{aligned} \varsigma_{\sigma ti,\sigma tt,\rho,r} &\sim Gamma(.1,.1)\\ \nu_{\mu ss,\mu tt,\gamma} &\sim Gamma(.1,.1)\\ \varsigma_{\rho effect} &\sim Gamma(1,1) \end{aligned}$$

One should note that for several of the Osth and Dennis (2015) parameters, uninformative prior distributions were only employed in Experiment 1 and were otherwise specified using informed priors via kernel density estimation. Density estimation was conducted for a gridsize of 10,000 points on a region that comprised the entire posterior distribution plus one standard deviation above and below. Likelihoods for proposals that were between the grid points of the KDE were estimated using linear interpolation between the two nearest points. For each of the restricted models in the model selection exercise (the $\sigma_{ti}^2 = 0$,$\gamma = 0$, and z/a_{slope},a_{slope} models), the corresponding model's posteriors from the fit to Experiment 1 was used to generate the informed priors.

For all models, the number of chains was three times the number of parameters. After 20,000 burn-in iterations were discarded, the MCMC chains were thinned by only accepting 1 sample every 20th iteration. The process continued until 1500 MCMC samples were accepted for each chain. Convergence was assessed using the Gelman-Rubin statistic; a model was considered converged if this statistic was less than 1.1 for all parameters. This criterion was satisfied for all models. Chains were also visually assessed for convergence.

B.1. Posterior predictive simulations

To generate posterior predictive distributions, we selected one in every 30 samples from each participant's posterior distributions and used them to simulate model predictions. Simulations were of the same size as the original dataset.

B.2. Data exclusions

B.2.1. Experiment 1

All responses faster than .25 s and slower than 3 s were excluded prior to model fitting. In addition, the first test trial was found to have considerably longer RT ($M = 1.87$) than the grand mean ($M = .94$). For this reason, we additionally omitted data from the first trial. Both of these exclusions resulted in an omission of 2.8% of responses.

B.2.2. Criss (2010, Experiment 2)

All responses faster than .2 or slower than 2.5 s were removed. The first test trial from each block was removed due to higher RT ($M = .94$ s) than the grand mean ($M = .73$ s). This resulted in the exclusion of 1.3% of all responses.

B.2.3. Starns (2014, Experiment 2)

All responses faster than .2 or slower than 2.5 s were removed. The first test trial from each block was removed due to higher RT ($M = .96$ s) than the grand mean ($M = .78$ s). This resulted in the exclusion of 1.3% of all responses.

B.2.4. Annis et al. (2016, Experiment 2)

Due to some very long RTs in the LD condition, we adopted somewhat different exclusion criteria for this dataset. Prior to applying the models, all responses faster than .25 or slower than 10 s were removed. Three participants were excluded for having over 20% of their RTs removed under such criteria. Subsequently, we removed all RTs that were greater than 3 standard deviations above a participant's mean RT. The first test trial from each block was removed due to higher RT ($M = 3.25$ s) than the grand mean ($M = 1.40$ s). This resulted in the exclusion of 3.5% of all responses.

Appendix C. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cogpsych.2018.04.002.

References

Annis, J., Dube, C., & Malmberg, K. J. (2016). A Bayesian approach to discriminating biased responding and sequential dependencies in binary choice data. *Decision*.

Annis, J., Malmberg, K. J., Criss, A. H., & Shiffrin, R. M. (2013). Sources of interference in recognition testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(5), 1365–1376.

Averell, L., Prince, M., & Heathcote, A. (2016). Fundamental causes of systematic and random variability in recognition memory. *Journal of Memory and Language, 88*, 51–69.

Bowen, H. J., Spaniol, J., Patel, R., & Voss, A. (2016). A diffusion model analysis of decision biases affecting delayed recognition of emotional stimuli. *PLoS ONE, 11*(1).

Bower, G. H. (1972). Stimulus sampling theory of encoding variability. In A. W. Melton, & E. Martin (Eds.). *Coding processes in human memory* (pp. 85–121). Wiley.

Brockdorff, N., & Lamberts, K. (2000). A feature-sampling account of the time course of old-new recognition judgments. *Journal of Experimental Psychology-Learning Memory and Cognition, 26*(1), 77–102.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57*, 153–178.

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language, 49*(2), 231–248.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin and Review, 3*(1), 37–60.

Cox, G. E., & Shiffrin, R. M. (2012). Criterion setting and the dynamics of recognition memory. *Topics in Cognitive Science, 4*, 135–150.

Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language, 55*, 461–478.

Criss, A. H. (2010). Differentiation and response bias in episodic memory: Evidence from reaction time distributions. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 36*(2), 484–499.

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

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H., & Usher, M. (2005). The demise of short-term memory revisited: Empirical and computational investigations of recency effects. *Psychological Review, 112*(1), 3–42.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108*(2), 452–478.

Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language, 59*, 361–376.

Divis, K. M., & Benjamin, A. S. (2014). Retrieval speeds context fluctuation: Why semantic generation enhances later learning but hinders prior learning. *Memory & Cognition, 42*, 1049–1062.

Donkin, C., & Nosofsky, R. M. (2012a). A power-law model of psychological memory strength in short- and long-term recognition. *Psychological Science, 23*(6), 625–634.

Donkin, C., & Nosofsky, R. M. (2012b). The structure of short-term memory scanning: An investigation using response time distribution models. *Psychonomic Bulletin & Review, 19*, 363–394.

Dosher, B. A. (1981). The effects of delay and interference: A speed-accuracy study. *Cognitive Psychology, 13*(4), 551–582.

Dosher, B. A. (1984). Degree of learning and retrieval speed: Study time and multiple exposures. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(4), 541–574.

Dutilh, G., Vandekerckhove, J., Forstmann, B. U., Keuleeurs, E., Brysbaert, M., & Wagenmakers, E. J. (2012). Testing theories of post-error slowing. *Attention, Perception, & Psychophysics, 74*(2), 454–465.

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review, 62*, 145–154.

Gehring, R. E., Toglia, M. P., & Kimble, G. A. (1976). Recognition memory for words and pictures at short and long retention intervals. *Memory & Cognition, 4*(3), 256–260.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*(1), 1–67.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory and Cognition, 13*(1), 8–20.

Glanzer, M., Hilford, A., & Maloney, L. T. (2009). Likelihood ratio decisions in memory: Three implied regularities. *Psychonomic Bulletin & Review, 16*(3), 431–455.

Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior, 15*, 1–16.

Glenberg, A. M., & Kraus, T. A. (1981). Long-term recency is not found on a recognition test. *Journal of Experimental Psychology: Human Learning and Memory, 7*, 475–479.

Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(6), 1210–1230.

Hintzman, D. L., Caulton, D. A., & Curran, T. (1994). Retrieval constraints and the mirror effect. *Journal of Experimental Psychology-Learning Memory and Cognition, 20*(2), 275–289.

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments – Evidence for separate processes of familiarity and recall. *Journal of Memory and Language, 33*(1), 1–18.

Howard, M. W. (2014). Mathematical learning theory through time. *Journal of Mathematical Psychology, 59*, 18–29.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 923–941.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*, 268–299.

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review, 96*(2), 208–233.

Humphreys, M. S., Pike, R., Bain, J. D., & Tehan, G. (1989). Global matching: A comparison of the SAM, Minerva II, Matrix and TODAM models. *Journal of Mathematical Psychology, 33*, 36–67.

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review, 12*(5), 852–857.

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

Jones, M., & Dzhafarov, E. N. (2014). Unfalsifiability and mutual translatibility of major modeling schemes for choice reaction time. *Psychological Review, 121*(1), 1–32.

Jonker, T. R., Seli, P., & MacLeod, C. M. (2013). Putting retrieval-induced forgetting in context: An inhibition-free, context-based account. *Psychological Review, 120*(4), 852–872.

Jou, J., Flores, S., Cortes, H. M., & Leka, B. G. (2016). The effects of weak versus strong relational judgments on response bias in Two-Alternative-Forced-Choice recognition: Is the test criterion-free? *Acta Psychologica, 167*, 30–44.

Karyanadis, F., Mansfield, E. L., Galloway, K. L., Smith, J. L., Provost, A., & Heathcote, A. (2009). Anticipatory reconfiguration elicited by fully and partially informative cues that validly predict a switch in task. *Cognitive, Affective, & Behavioral Neuroscience, 9*(2), 202–215.

Kary, A., Taylor, R., & Donkin, C. (2016). Using Bayes factors to test the predictions of models: A case study in visual working memory. *Journal of Mathematical Psychology, 72*, 210–219.

Kiliç, A., Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2017). Models that allow us to perceive the world more accurately also allow us to remember past events more accurately via differentiation. *Cognitive Psychology, 92*, 65–86.

Kiliç, A., & Öztekin, I. (2014). Retrieval dynamics of the strength based mirror effect in recognition memory. *Journal of Memory and Language, 76*, 158–173.

Kinnell, A., & Dennis, S. (2011). The list length effect in recognition memory: An analysis of potential confounds. *Memory & Cognition, 39*, 348–363.

Kinnell, A., & Dennis, S. (2012). The role of stimulus type in list length effects in recognition memory. *Memory & Cognition, 40*, 311–325.

Klein, K. A., Shiffrin, R. M., & Criss, A. H. (2007). Putting context in context. *The foundations of remembering: Essays in honor of Henry L. Roediger III* (pp. 171–189). Psychology Press.

Koop, G., Criss, A. H., & Malmberg, K. J. (2015). The role of mnemonic processes in pure-target and pure-foil recognition memory. *Psychonomic Bulletin & Review, 22*(2), 509–516.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology, 55*, 1–7.

Lee, M. D., & Dry, M. J. (2006). Decision making and confidence given uncertain advice. *Cognitive Science, 30*, 1081–1095.

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2015). Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. *Psychological Review, 122*(2), 337–363.

Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology, 57*, 335–384.

Malmberg, K. J., & Annis, J. (2012). On the relationship between memory and perception: Sequential dependencies in recognition memory testing. *Journal of Experimental Psychology: General, 141*(2), 233–259.

Malmberg, K. J., Criss, A. H., Gangwani, T. H., & Shiffrin, R. M. (2012). Overcoming the negative consequences of interference from recognition memory testing. *Psychological Science, 23*(2), 115–119.

Marr, D. (1971). A theory for archicortex. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 262*(841), 23–81.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*(4), 724–760.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*(3), 419–457.

McElree, B., & Dosher, A. (1989). Serial position and set size in short-term memory: The time course of recognition. *Journal of Experimental Psychology: General, 118*, 346–373.

McKoon, G., & Ratcliff, R. (1979). Priming in episodic and semantic memory. *Journal of Verbal Learning and Verbal Behavior, 18*(4), 463–480.

Mensink, G. J., & Raaijmakers, J. G. W. (1988). A model for interference and forgetting. *Psychological Review, 95*(4), 434–455.

Murdock, B. B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review, 104*(4), 839–862.

Murdock, B. B., & Anderson, R. E. (1975). Encoding, storage, and retrieval of item information. In R. L. Solso (Ed.). *Theories in cognitive psychology: The Loyola Symposium*. Erlbaum.

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.

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review, 118*(2), 280–315.

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

Osth, A. F., & Dennis, S. (2014). Associative recognition and the list strength paradigm. *Memory & Cognition, 42*(4), 583–594.

Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review, 122*(2), 260–311.

Osth, A. F., Dennis, S., & Heathcote, A. (2017). Likelihood ratio sequential sampling models of recognition memory. *Cognitive Psychology, 92*, 101–126.

Pastötter, B., Schicker, S., Niedernhuber, J., & Bäuml, K. T. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(2), 287–297.

Peixotto, H. E. (1947). Proactive inhibition in the recognition of nonsense syllables. *Journal of Experimental Psychology, 37*(1), 81–91.

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

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59–108.

Ratcliff, R. (2006). Modeling response signal and response time data. *Cognitive Psychology, 3*(53), 195–237.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*, 873–922.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9*(5), 347–356.

Ratcliff, R., Sheu, C. F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99*(3), 518–535.

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116*(1), 59–83.

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review, 120*(3), 697–719.

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*, 438–481.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106*(2), 261–300.

Reed, A. V. (1973). Speed-accuracy trade-off in recognition memory. *Science, 181*(4099), 574–576.

Reed, A. V. (1976). List length and the time course of recognition in immediate memory. *Memory & Cognition, 4*(1), 16–30.

Rotello, C. M., & Heit, E. (1999). Two-process models of recognition memory: Evidence for recall-to-reject? *Journal of Memory and Language, 40*(3), 432–453.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12*(4), 573–604.

Sahakyan, L., & Hendricks, H. E. (2012). Context change and retrieval difficulty in the list-before-last paradigm. *Memory & Cognition, 40*(6), 844–860.

Sahakyan, L., & Smith, J. R. (2014). "A long time ago, in a context far, far away": Retrospective time estimates and internal context change. *Journal of Experimental Psychology: General, 40*, 86–93.

Schmitz, F., & Voss, A. (2012). Decomposing task-switching costs with the diffusion model. *Journal of Experimental Psychology: Human Perception and Performance, 38*(1), 222–250.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels.* MIT Press.

Schulman, A. L. (1974). The declining course of recognition memory. *Memory and Cognition, 2*, 14–18.

Schwartz, G., Howard, M. W., Jing, B., & Kahana, M. J. (2005). Shadows of the Past: Temporal retrieval effects in recognition memory. *Psychological Science, 16*(11), 898–904.

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review, 115*(4), 893–912.

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

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

Siegel, L. L., & Kahana, M. J. (2014). A retrieved context account of spacing and repetition effects in free recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 40*(3), 755–764.

Starns, J. J. (2014). Using response time modeling to distinguish memory and decision processes in recognition and source tasks. *Memory & Cognition, 42*, 1357–1372.

Starns, J. J., Chen, T., & Staub, A. (2017). Eye movements in forced-choice recognition: Absolute judgments can preclude relative judgments. *Journal of Memory and Language, 93*, 55–66.

Starns, J. J., & Ratcliff, R. (2014). Validating the unequal-variance assumption in recognition memory using response time distributions instead of ROC functions: A diffusion model analysis. *Journal of Memory and Language, 70*, 36–52.

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of the zROC slopes with response time data and the diffusion model. *Cognitive Psychology, 64*, 1–34.

Starns, J. J., Ratcliff, R., & White, C. N. (2012). Diffusion model drift rates can be influenced by decision processes: An analysis of the strength-based mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(5), 1137–1151.

Starns, J. J., White, C. N., & Ratcliff, R. (2010). A direct test of the differentiation mechanism: REM, BCDMEM, and the strength-based mirror effect in recognition memory. *Journal of Memory and Language, 63*, 18–34.

Sternberg, S. (1966). High-speed scanning in human memory. *Science, 153*(3736), 652–654.

Strong, E. K. J. (1912). The effect of length of series upon recognition memory. *Psychological Review, 19*, 447–462.

Talmi, D., & Goshen-Gottstein, Y. (2006). The long-term recency effect in recognition memory. *Memory, 14*(4), 424–436.

Turner, B. M., Dennis, S., & Van Zandt, T. (2013). Likelihood-free Bayesian analysis of memory models. *Psychological Review, 120*(3), 667–678.

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods, 18*(3), 368–384.

Underwood, B. J. (1978). Recognition memory as a function of the length of study list. *Bulletin of the Psychonomic Society, 12*, 89–91.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*, 550–592.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods, 16*(1), 44–62.

Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review, 19*(6), 1047–1056.

Vickers, D. (1979). *Decision processes in visual perception.* Academic Press.

Vickers, D., & Lee, M. D. (1998). Dynamic models of simple judgments: I. Properties of a self-regulating accumulator module. *Non-Linear Dynamics, Psychology, and Life Sciences, 2*, 169–194.

Wagenmakers, E. J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology, 21*, 641–671.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research, 11*, 3571–3594.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*(1), 152–176.