Modeling the dynamics of recognition memory testing

Adam F. Osth
University of Melbourne

Anna Jansson
University of Newcastle

Simon Dennis
University of Newcastle

Andrew Heathcote
University of Tasmania

Address correspondence to:
Adam Osth
E-mail: adamosth@gmail.com
Abstract

A robust finding in recognition memory is the observation that performance declines monotonically across test trials. Despite the prevalence of this result, there is a lack of consensus on the mechanism responsible for the decline. Three hypotheses have been put forward: 1.) interference is caused by the 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 in the literature on testing effects in recognition memory: 1.) the finding that the number of test items plays a larger role in determining performance than the number of studied items, 2.) the finding that performance decreases less for strong items than weak items in pure lists but not in mixed lists, and 3.) the finding that lexical decision trials interspersed between recognition test trials do not impact performance. Analysis of the resulting parameter estimates suggests that item interference plays a very weak role in explaining the effects of recognition testing, while context drift plays a very large role. Changes in speed-accuracy thresholds did not show a consistent tendency either upward or downward, but were demonstrated to show a very large role in moderating the size of the testing effect. These results are consistent with prior work showing a weak role for item noise in recognition memory and other results demonstrating that retrieval is a strong cause of context change in episodic memory.
Modeling the dynamics of recognition memory testing

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 (Nosofsky, Little, Donkin, & Fific, 2011; Sternberg, 1966; McElree & Dosher, 1989) and long (Clark & Gronlund, 1996; Dennis & Humphreys, 2001; Gillund & Shiffrin, 1984; McCleland & 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.

A nearly universal finding in recognition memory research is that recognition memory performance decreases through the course of testing, with later test trials performing more poorly than earlier test trials. This finding was first reported by Peixotto (1947) but has been replicated extensively in the many decades since its original publication (Annis, Malmberg, Criss, & Shiffrin, 2013; Averell, Prince, & Heathcote, 2016; Criss, Malmberg, & Shiffrin, 2011; Kiliç, Criss, Malmberg, & Shiffrin, 2017; Kim & Glanzer, 1995; Malmberg, Criss, Gangwani, & Shiffrin, 2012; Murdock & Anderson, 1975; Ratcliff & Hockley, 1980; Schulman, 1974; Shiffrin, Huber, & Marinelli, 1995). However, despite its status as an empirical regularity, theoretical interest in the nature of the testing effect has emerged somewhat more recently (e.g.: Criss et al., 2011; Osth & Dennis, 2015). The finding has been referred to as "output interference" in some parts of the literature. We will instead be referring to the finding as the test position effect, or TPE, to avoid commitment to the idea that the effect is driven by interference. Modeling of the TPE has explored causal factors acting through the decision process or memory retrieval, but not both. This work aims to bridge that gap by introducing a new combined model of memory retrieval and decision making.
Causes of the TPE

The earliest known attempt to explore the nature of the TPE came from Ratcliff (1978)’s attempt to model the phenomenon with his Diffusion Decision Model (DDM), an evidence accumulation model of the decision process. In the DDM, evidence begins to accumulate at the starting point $z$ in a noisy fashion toward one of two response boundaries; an upper response boundary denoted by the parameter $a$ corresponding to a 'YES' decision, and a lower bound at zero corresponding to a 'NO' decision. The boundary that is reached determines the choice made by the participant while the time taken to reach the boundary is the response time (RT). The speed-accuracy tradeoff is captured with 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, and is denoted by $v$. Increases in the absolute value of $v$ increase the proportion of correct responses and decrease the RT. Drift rate is not fixed within a trial but varies from trial-to-trial with standard deviation $\eta$, which is analogous to cross-trial variability in memory strength in signal detection theory (SDT). Finally, nondecision time components such as perceptual processing and response output are modeled by parameter $t_{ER}$ as a shift to the RT distribution. Variability in nondecision time was later introduced by Ratcliff, Gomez, and McKoon (2004) as a uniform distribution with width $s_t$. A diagram of the DDM is depicted in Figure 1.

One might naively assume that changes in performance are due to changes in memory strength alone. An advantage of the DDM is its ability to fit both choice and response time to yield separate estimates of memory strength (as measured by the drift rate) and the speed-accuracy threshold, a feat that is not possible when only choice data are considered. Drift rate and speed-accuracy thresholds can be separably estimated 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
Figure 1. Basic description of the Ratcliff diffusion model (DDM). The drift rate is a sample from a normal distribution with mean $v$ and standard deviation $\eta$ relative to a drift criterion $d_c$. This drift rate is then used to drive a diffusion process (bottom panel). During the diffusion process, evidence accumulates through the trial, beginning at the starting point $z$, and continues until either the upper response boundary ($a$) or the lower response boundary (0) is reached. The boundary that is reached is the choice whereas the time taken to reach the boundary is the response time (RT). Evidence accumulation is noisy, such that diffusion processes with the same drift rate will often reach different boundaries and produce different RTs. Depicted are three sample trajectories with the same drift rate.

Leading edge and the skew of the RT distribution (Ratcliff & McKoon, 2008). Ratcliff (1978) reasoned that drift rates, bias, and response boundaries could all be affected during the course of testing, as participants could compensate for their decreases in accuracy by changing their speed-accuracy threshold. Results of the modeling supported such a conjecture, with decreases in drift rate being accompanied by increases in response boundaries and decreases in response bias. These results suggest that recognition memory testing affects both memory strength as well as decision level components.

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) emphasized that there may be multiple sources of the TPE within the context of the search of associative memory (SAM) model, a global matching model of recognition memory. In global matching models (see Clark & Gronlund, 1996, for a review) studied items are bound to a representation of the study episode, what is referred to as the 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. What was considered a major strength of the framework at the time was the natural ability of such models to capture the list length effect, the finding that performance is worse after studying a longer list (Strong, 1912); as the size of the memory set is increased, the variance of the familiarity distributions increases and the signal-to-noise ratio decreases.

Gillund and Shiffrin reasoned that a simple explanation of the TPE is that test items are added to memory, increasing interference through the course of testing. 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. This is an account we refer to as the item noise account of the TPE. The idea that test items are added into memory is an extremely reasonable assumption, especially given that studies have demonstrated that participants have memory for the lures tested in an experiment (e.g.: Jacoby, Shimizu, Daniels, & Rhodes, 2005). Nonetheless, Gillund and Shiffrin also acknowledged that an additional possibility was that the increasing retention interval through the course of testing, which could be
modeled as a decreasing match between the context cue employed at test and the stored context from the study episode.

At the time, Gillund and Shiffrin were not able to distinguish between these two possibilities. Nonetheless, Criss et al. (2011) revisited the issue and presented evidence against the retention interval account of the TPE. In particular, Criss et al. manipulated the retention interval in that participants were either tested immediately after completion of the study list or after a 20 minute filled delay. Performance was worse in the delayed condition, but contrary to the retention interval account of the TPE, the decreases in performance through testing were parallel to the immediate testing condition. More critically, 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, a result that severely undermined the retention interval account. Criss et al. were able to successfully model the decreases in performance with the retrieving effectively from memory (REM) model (Shiffrin & Steyvers, 1997), a model which shares many assumptions with the SAM model, including global matching and the assumption that forgetting is primarily due to interference among the items. 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.

Evidence Against the Item Noise Account of the TPE

Among the first inconsistencies in the item noise account of the TPE is that recent investigations of the list length effect have found that the effect 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 artifically
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), decreasing 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 on the beginning items. When these confounds were controlled, investigations resulted in either no effect (Dennis & Humphreys, 2001; Dennis, Lee, & Kinnell, 2008; Kinnell & Dennis, 2011, 2012; Schulman, 1974) or a small effect (Cary & Reder, 2003; Kinnell & Dennis, 2012, for select non-linguistic stimuli) on recognition memory performance. An additional possibility that we will consider in this article is the possibility that participants may adopt different speed-accuracy thresholds across the list length conditions (e.g.: Donkin & Nosofsky, 2012b).

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

Osth and Dennis (2015) argued against item noise accounts of forgetting on the basis of modeling with a model of recognition memory based on the matrix model by Humphreys, Bain, and Pike (1989). In the matrix model, both items and contexts are represented as vectors and item-context bindings are represented as outer products. At study, items and contexts are bound together and summed into a memory matrix that is a composite of all of the studied memories. At test, the probe item cue and the context cue
are bound together and matched against the memory matrix. Given that the memory matrix is a composite of stored memories, this is mathematically equivalent to matching against each of the memories separately, like is done with the SAM model. In some cases, the models have even been found to make identical predictions (Humphreys, Pike, Bain, & Tehan, 1989).

The Osth and Dennis model differed from the original Humphreys et al. model by parameterizing the similarities between the cue vectors and the vectors stored in memory. As the noise in the similarity between dissimilar items is increased, item noise increases in the model as a consequence of the confusability between the items. However, as the noise between the context cue and previously stored contexts was increased, context noise and background noise result from confusions with item-context bindings acquired prior to the study episode. Osth and Dennis comprehensively applied the model to ten recognition memory datasets with 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.

While Osth and Dennis estimated a very low amount of item noise, they did not model the TPE in their modeling work; all item noise was assumed to be generated by matching to memories acquired during the study phase. Nonetheless, they simulated the test lists of the Schulman dataset using the estimated parameters from their fits along with an assumption that test items incremented item noise; this simulation indicated that their estimated magnitudes of item noise failed to generate enough interference to capture the TPE. They were, however, able to give a very 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 cue
memory and pushes it further away from the context of the study episode, 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.

Context drift has not generally been employed in recent models of recognition memory (but see Murdock, 1997, for an exception) but has a long history in models of episodic memory more generally (see Howard, 2014, for a review). Estes initially used stimulus sampling theory as a model of spontaneous recovery; after a sequence of trials where a conditioned response is unlearned, the context probabilistically changes to the point where it may no longer match the unlearning trials but matches the conditions of initial learning, leading to recovery of the conditioned response. Context drift was introduced into episodic memory models by Bower (1972) and was later used by Glenberg (1976) to unify the spacing effect and the recency effect; 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.

Context drift is a relatively simple mechanism and has been a major feature in models of free recall (Davelaar, Goshen-Gottstein, Ashkenazi, Haarmann, & Usher, 2005; Mensink & Raaijmakers, 1988; Polyn, Norman, & Kahana, 2009; Sederberg, Howard, & Kahana, 2008) due its power and simplicity. In existing models it is 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. The 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.

**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 response times 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 boundaries, 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 variation 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 of memory strength as 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 we will discuss, the Osth and Dennis model has the ability to parameterize the degree of item noise and implement assumptions regarding context drift. Thus, measurement of both of these parameters can be used to get an understanding of which parameter is more responsible for the performance decrements caused by recognition memory testing. We additionally measure changes in response bias and boundaries by implementing linear functions of each of these parameters over test trials. While prior modeling work measured testing effects by fitting to blocks of recognition testing data (Criss et al., 2011; Kiliç et al., 2017; Ratcliff,
DYNAMICS OF RECOGNITION TESTING

1978), we fit the model to each test trial of the test sequence. 

The Osth and Dennis model is based on the matrix model (Pike, 1984; Humphreys, Bain, & Pike, 1989). The fundamental constituents of the 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:

\[
M = \sum rC_s \otimes I_t
\]

where \(r\) is a learning rate parameter. The subscript \(s\) indicates the context vector corresponds to the study episode and subscript \(t\) indicates the item vector is a target. 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 \(M\):

\[
s = (C'_s \otimes I'_t).M
\]

where the dashes indicate that the cues employed may not completely resemble 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. Following Humphreys, Pike, et al. (1989), we can deconstruct Equation 2 into the various components that comprise \(M\), assuming that memory is being probed for a target item:
\[ s = (C_s' \otimes I_t') \cdot r(C_s \otimes I_t) + \sum_{i \in L, i \neq t} r(C_s \otimes I_i) + \sum_{u \in P, u \neq s} (C_u \otimes I_t) + \sum_{u \in P, u \neq s, z \in L} (C_u \otimes I_z) \]

The right hand column of the first row is the **self match** as the stored vectors are the original studied item \( I_t \) in the study context \( C_s \). The self match is the primary determinant of discrimination as it is not present in the calculation of memory strength for lures.

The **item noise** term comprises items that are not the target item (the subscript \( i \) indicates the item is not the target and \( L \) is the set of study list items) but were nonetheless present in the study list context and thus learned with learning rate \( r \). Similarity between the item cue \( I_t' \) and the other items on the list \( I_i \) increases the degree of item noise at retrieval.

The **context noise** term comprises prior occurrences of the item cue; there is a match to the item cue \( I_t' \), but the subscript \( u \) in the context vector refers to a context in the set \( P \) that refers to all contexts experienced prior to the study list context. Similarity between the context cue \( C_s' \) and the stored contexts \( C_u \) increases the degree of context noise.

The final term comprises the **background noise**, which are items that were not present on the study list and were also not present in the study list context. Similarity between the test cues and these memories increases the overall degree of interference. Background noise has a history of incorporation into the TODAM (Murdock & Kahana, 1993a, 1993b; Murdock, 1997) and SAM models (Gronlund & Elam, 1994). Equation 3 can then be rewritten as the match between the cue vectors and the stored vectors:
\[ s = r(C'_s, C_s)(I'_t, I_t) + \sum_{i \in L, i \neq t} r(C'_s, C_s)(I'_t, I_t) + \sum_{u \in P, u \neq s} (C'_s, C_u)(I'_t, I_t) + \sum_{u \in P, u \neq s, z \in L} (C'_s, C_u)(I'_t, I_i) \]

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

\[ C'_s, C_s \sim \text{Normal}(\mu_{ss}, 0) \quad \text{Context Match} \quad (5) \]
\[ C'_s, C_u \sim \text{Normal}(0, \sigma^2_{su}) \quad \text{Context Mismatch} \]
\[ I'_t, I_t \sim \text{Normal}(\mu_{tt}, \sigma^2_{tt}) \quad \text{Item Match} \]
\[ I'_t, I_i \sim \text{Normal}(0, \sigma^2_{ti}) \quad \text{Item Mismatch} \]

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 avoids covariances during the context drift process.

The distributions of the matches and mismatches from Equation 5 are substituted into the terms for Equation 4 to derive mean and variance expressions for the signal and noise distributions. Because each interference term is the multiplication of an item match/mismatch by a context match/mismatch, and each are represented by normal
distributions, 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 and new distributions are as follows:

\[
\mu_{\text{old}} = r\mu_{ss}\mu_{tt}
\]
\[
\mu_{\text{new}} = 0
\]

\[
\sigma_{\text{old}}^2 = r^2(\mu_{ss}^2\sigma_{tt}^2) + r^2(l - 1)(\mu_{ss}^2\sigma_{ti}^2) + m(\mu_{tt}^2\sigma_{su}^2 + \sigma_{su}^2\sigma_{ti}^2) + n(\sigma_{su}^2\sigma_{ti}^2)
\]

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 Equation 7 can be viewed as the contributions of the self match, item noise, context noise, and
background noise. Equations 7 and 8 are identical with the exception of the self match variance term which is only in Equation 7 and the fact that item noise is scaled by $l - 1$ in Equation 8 instead of $l$. We also can further simplify the model by combining the $m$ and $\sigma^2_{su}$ terms into a the parameter $\rho$ 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 $\beta$. The simplified variance equations are as follows:

\[
\sigma_{old}^2 = r^2(\mu_{ss}^2 \sigma_{ti}^2) + \text{Self Match} \\
\quad + r^2(l - 1)(\mu_{ss}^2 \sigma_{ti}^2 + \sigma_{ss}^2 \sigma_{ti}^2) + \text{Item Noise} \\
\quad + (\mu_{tt}^2 + \rho \sigma_{ti}^2) + \text{Context Noise} \\
\quad + \beta \text{ Background Noise}
\]

\[
\sigma_{new}^2 = r^2 l(\mu_{ss}^2 \sigma_{ti}^2) + \text{Item Noise} \\
\quad + (\mu_{tt}^2 + \rho \sigma_{ti}^2) + \text{Context Noise} \\
\quad + \beta \text{ Background Noise}
\]

The roles of each of the parameters were described in detail by Osth and Dennis (2015). Nonetheless, we reproduce description for the crucially relevant parameters here. Each interference term in Equations 9 and 10 arises from combinations of the matches and mismatches to the context and item cues. The mean of the target distribution is a multiplication of the learning rate $r$, the mean of the context match $\mu_{ss}$, and the mean of the item match $\mu_{tt}$. $\mu_{ss}$ serves two direct roles in our fits to data, a.) it is expected to vary with retention interval, as an increased retention interval could result in a deficit in reinstating the original study context representation and b.) it is expected to decrease during test trials if context drift is implemented. Each of these consequently reduces the
mean of the target distribution. The mean of the item match $\mu_{tt}$ can be conceptualized as the strength of an item cue (e.g.: Eich, 1982). However, in virtually all of the applications we will fix $\mu_{tt}$ to one, with the exception of a task switching manipulation where we use lower values to a potential deficit in employing the probe item cue.

The self match variance is primarily determined by the item match variability parameter $\sigma^2_{tt}$. The parameter psychologically 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 is what differentiates the variability of the target distribution from that of the lure distribution. Thus, if $\sigma^2_{tt}$ is greater than zero, there is greater variability in the target distribution than the lure distribution, allowing the model to be consistent with findings from receiver operating characteristics (ROCs: Heathcote, 2003; Ratcliff, Sheu, & Gronlund, 1992; Wixted, 2007). Recent evidence using the DDM has generalized this finding to RT distributions; several investigations have demonstrated that drift rate distributions have a larger standard deviation for targets than for lures (Osth, Dennis, & Heathcote, 2017; Osth, Bora, Dennis, & Heathcote, 2017; Starns, Ratcliff, & McKoon, 2012; Starns & Ratcliff, 2014; Starns, 2014). We will demonstrate later that the estimated parameters were sufficient to produce estimates of target-to-lure variability that are comparable in magnitude to previous explorations with the DDM.

The item noise in Equations 9 and 10 is crucially determined by the item mismatch variability parameter $\sigma^2_{ti}$. If $\sigma^2_{ti} = 0$, the entire item noise term reduces to zero. If $\sigma^2_{ti}$ is positive, item noise is increased by a.) the number of items on the list $l$, b.) the learning rate of the studied items $r$, and c.) the match to the study context $\mu_{ss}$. Increased item noise through the course of testing can be implemented by including additional item noise terms for each test trial. This is not the same as incrementing $l$ because the test items likely have a different strength of match to the current context than the studied items. Thus, for each test item, a separate item noise term is added with item mismatch
variability $\sigma^2_{ti}$ but are stored with a higher value of $\mu_{ss}$ due to their greater recency than the study list items. As we will discuss below, this leads to some interesting predictions from the model as it is possible for the model to predict a TPE that is larger than the list length effect. Osth and Dennis (2015) allowed the $\sigma^2_{ti}$ parameter to vary across stimulus classes to capture the different degrees of item noise elicited by each stimulus class (e.g., Kinnell & Dennis, 2012; Osth, Dennis, & Kinnell, 2014). However, all of our applications of the present model use words as stimuli and thus a single value of $\sigma^2_{ti}$ is estimated for each participant that does not differ across conditions.

The context noise term is primarily determined by the context mismatch variability parameter $\rho$. $\rho$ is expected to increase with prior occurrences of the cue. In our applications, this is manipulated through normative word frequency.

Background noise cannot be manipulated, but its presence produces a constant source of interference that is unaffected by experimental manipulations. In our applications, we simplified the number of parameters by fixing $\beta$ to .05, which was approximately the mean of the background noise contribution in previous fits Osth and Dennis (2015).

The Likelihood Ratio Transformation

One may notice that manipulations of parameters such as $\rho$ and $\sigma^2_{ti}$ only serve to increase the variances of the memory strength distributions and have no effect on $\mu_{old}$. Osth and Dennis (2015) demonstrated that changes in the variances of the underlying distributions was insufficient to produce the full range of mirror effects demonstrated in 2AFC performance. The mirror effect is the 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 $\mu_{LFnew} < \mu_{HFnew} < \mu_{HFold} < \mu_{LFold}$.

Such a distributional arrangement can be obtained via a log likelihood ratio
transformation of the memory strengths, where the log likelihood ratio $\lambda$ of a strength value $x$ is $\lambda(x) = \log \left[ \frac{f_{\text{old}}(x)}{f_{\text{new}}(x)} \right]$. When the strength distributions of $x$ contain unequal variance, the distribution of $\lambda$ is noncentral chisquare in shape (Glanzer, Hilford, & Maloney, 2009). We instead employ the linear approximation devised by Osth, Dennis, and Heathcote (2017) which results in normal distributions, making the model congruent with the analytics of the DDM which are expressed for the normal case. 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 $\sigma_{\text{new}}$:

$$
d = \frac{\mu_{\text{old}}}{\sigma_{\text{new}}} \quad (11)
$$

$$
S = \frac{\sigma_{\text{old}}}{\sigma_{\text{new}}} \quad (12)
$$

The means and standard deviations of $\lambda$ can be expressed in terms of $d$ and $S$ via a linear approximation of the log likelihood ratio transformation (details of which can be found in Osth, Dennis, & Heathcote, 2017) resulting in normal distributions with the following means and standard deviations:

$$
\mu_{\lambda_{\text{new}}} = -((d^2/2)(S^2/4S^2) + \log(S)) \quad (13)
$$

$$
\mu_{\lambda_{\text{old}}} = d(S^2/2S^2) + \mu_{\lambda_{\text{new}}} \quad (14)
$$

$$
\sigma_{\lambda_{\text{new}}} = d(S^2/2S^2) \quad (15)
$$

$$
\sigma_{\lambda_{\text{old}}} = S\sigma_{\lambda L} \quad (16)
$$

Inspection of Equations 15 and 16 reveals that the standard deviations of $\lambda$ increase linearly with $d$. This is quite contrary to standard conventions of the DDM, where drift rate variability is often fixed across several conditions in an experiment. Nonetheless, Osth, Dennis, and Heathcote (2017) developed a version of the DDM with a log likelihood ratio DDM where Equations 13 and 14 specified the mean drift rate for lures and targets and
Equations 15 and 16 specified the standard deviations. The model proved to be quite capable of addressing benchmark RT distributions across several datasets and was also found to produce nearly identical drift rates to those estimated from the standard DDM. They also found that the model benefited from the addition of a drift criterion $d_c$, which gets subtracted from Equations 13 and 14. For this reason, we additionally fit $d_c$ as a free parameter for each participant but is fixed across all conditions within an experiment. This was partly motivated by the findings that there is substantial individual variation in response criteria across participants (Aminoff et al., 2012; Kantner & Lindsay, 2012).

Osth, Dennis, and Heathcote (2017) also demonstrated that the ratio of target-to-lure variability $S$ serves an additional role in the likelihood ratio model beyond the ROC function or differences in RT distributions between targets and lures. Specifically, when $S$ is greater than one, the mirror effect becomes asymmetric, with HR changing to a greater degree than the FAR as $d$ is increased. Fits of the log likelihood ratio diffusion model produced estimates of $S$ that were considerably greater than 1 across four datasets, which was not only consistent with the standard DDM's analysis of RT distributions, but also consistent with the asymmetric mirror effects in the data which where the changes in the HR were greater than the FAR for manipulations of strength and word frequency. The effect of test position is also consistent with this asymmetry, with HR changing to a greater degree than the FAR in both prior analyses (Criss et al., 2011; Kiliç et al., 2017) and in this article.

In conditions where participants study a mixed strength study list, such as when items were studied once or three times, if participants are not given a cue as to which strength they will be tested on then it becomes more appropriate for the expected strengths in the likelihood ratio transformation to be the average of the strong and weak items. 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. Usage of 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 can be seen below (lure expressions are unchanged):

$$
\mu_{old} = dd^* \frac{S^2 + 1}{2S^2} + \mu_{\lambda L} \quad (17)
$$

$$
\sigma_{old} = S^* \sigma_{\lambda L} \quad (18)
$$

**Context drift**

Implementation of context drift requires a function that relates context strength to the number of tested trials. 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 $i$:

$$
C_j = \gamma C_i + (\sqrt{1-\gamma^2})z \quad (19)
$$

where $\gamma$ 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 $Var[C_i.z] = 0^1$, $z$ no longer plays a role and we can express $\mu_{ss}$ on trial $j$ as:

$$
\mu_{ss,j} = \gamma \mu_{ss,i} \quad (20)
$$

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

$$
\mu_{ss,l} = \gamma^l \mu_{ss,i} \quad (21)
$$

---

1Such an assumption would be met if the two context vectors were orthogonal to each other (e.g.: Howard & Kahana, 2002).
where $\mu_{ss,i}$ behaves as an intercept that decays exponentially over successive trials. For all immediate testing conditions, $\mu_{ss}$ 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 used the free parameter $\mu_{ss,\text{delay}}$ 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.

Context drift occurs for both items that were stored at study along with 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 ($\mu_{ss} = 1$) and drift in subsequent trials. This means that when $\gamma < 1$, recent test items are more similar to the current test context than older test items. Given that $\mu_{ss}$ scales the item noise in Equations 9 and 10, this means that recent test trials exhibit greater item noise than older test trials. One should note that other assumptions may be possible. For instance, participants might make a distinction between the contexts of study and test, with test items being associated to the test context 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. 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.

In delayed testing conditions where testing begins with a partially reinstated context ($\mu_{ss,\text{delay}} < 1$ 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 $\gamma$ is estimated for each participant that does not vary across conditions.
Changes in Bias and Response Boundaries

Bias is represented in the diffusion model with the starting point $z$. Like with previous applications (Osth, Dennis, & Heathcote, 2017), we parameterize the starting point relative to the response boundary ($z/a$). We model changes in bias and boundaries with linear functions. Thus, unique values of $z/a$ and $a$ for trial $i$ can be predicted with only four parameters ($z/a_{\text{slope}}, z/a_{\text{intercept}}, a_{\text{slope}}, a_{\text{intercept}}$) with the following expressions:

\[
\frac{z}{a_i} = \frac{z}{a_{\text{slope}}}i + \frac{z}{a_{\text{intercept}}} \quad (22)
\]

\[
a_i = a_{\text{slope}}i + a_{\text{intercept}} \quad (23)
\]

Ratcliff (1978), in contrast, 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. While the choice of the linear function was an arbitrary decision, we found that they 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 (Ratcliff & McKoon, 2008; Donkin & Nosofsky, 2012b; Nosofsky et al., 2011). We follow suit here and vary both the slope and intercept parameters across manipulations of study-test delay and other cross-list manipulations, but hold the parameters fixed for within-list manipulations of stimulus difficulty.

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 matching front-end to model memory retrieval along with a back-end diffusion process for the decision stage. Through the course of this article, we applied the model to a number of datasets that are
theoretically challenging to existing accounts of the TPE. The first of which 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. The second and third datasets employ cross-list and between-list manipulations of strength, respectively, which we chose based on recent evidence showing that the TPE for strong and weak items depends on list composition (Kiliç et al., 2017). The final dataset employs a task-switching manipulation, which we chose based on evidence that lexical decision trials do not contribute to the TPE (Annis et al., 2013). A table documenting the number of participants, data points per participant, and conditions in this experiment and other experiments in the article can be found in Table 1.

Table 1

*Summary of the datasets fit by the model.*

<table>
<thead>
<tr>
<th>Dataset</th>
<th>N</th>
<th>Obs.</th>
<th>Cond.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td>107</td>
<td>233.48</td>
<td>List length (24 &amp; 96, cross-list), retention interval (immediate &amp; delayed, cross-list), WF (LF &amp; HF) Presentations (1x &amp; 5x, cross-list),</td>
</tr>
<tr>
<td>Criss (2010, E2)</td>
<td>16</td>
<td>1519.31</td>
<td>WF (LF &amp; HF, cross-list)</td>
</tr>
<tr>
<td>Starns (2014, E2)</td>
<td>33</td>
<td>290.24</td>
<td>Presentations (1x &amp; 3x)</td>
</tr>
<tr>
<td>Annis et al. (2016, E2)</td>
<td>59</td>
<td>308.89</td>
<td>Between-trial activity (blanks &amp; LD, cross-list)</td>
</tr>
</tbody>
</table>

*Notes: E = experiment, N = number of participants, Obs. = mean number of observations per participant, Cond. = conditions, LD = lexical decision, WF = word frequency, LF = low frequency, HF = high frequency*

Each of the datasets we employed use yes/no recognition testing. Criss et al. (2011) advocated the usage of two alternative forced choice tests (2AFC) on the basis of their observation of a criterion shift occurring in their experiment that utilized the yes/no testing
format. 2AFC tests have been assumed to be 'bias-free’ in that participants only make a relative judgment between the two items. If participants are making relative judgments only, investigations of the TPE can be safely assumed to not be contaminated by changes in bias occurring across test trials. However, recent investigations have undermined the idea that participants only make relative judgments in 2AFC testing. Jou, Flores, Cortes, and Leka (2016) found that response times were shorter for the left-most item in 2AFC displays and argued that participants likely terminate their decision on the left most item if they are confident it’s an old item. Starns, Chen, and Staub (2017) presented eyetracking evidence congruent with this speculation, demonstrating a substantial proportion of trials in which participants make a decision on the first seen item without even looking at the second item. If the absolute judgment on the first seen item is based on the extent to which that item’s memory strength exceeds a response criterion, then it becomes possible that changes in bias through testing additionally affect the 2AFC paradigm.

A larger concern introduced by these studies concerns how to conceptualize 2AFC testing in an evidence accumulation model. The DDM is an instance of a relative evidence model where the drift rate reflects the difference in the strength of evidence between the two alternatives. However, one could additionally construct a race model, where each alternative in a 2AFC trial receives its own accumulator with a drift rate that corresponds to its absolute strength (Brown & Heathcote, 2008; Usher & McClelland, 2001). These two types of models make distinct RT predictions on trials with two targets or two lures, a trial type referred to as null trials due to their lack of a correct answer, relative to trials with one target and one lure, a trial type referred to as valid trials. Relative evidence models predicting very slow RTs null trials due to the low differences between the drift rates of the two alternatives, while race models predict very fast RTs for trials with two targets due to the high drift rates of both alternatives and the slowest RTs for trials with two lures due to their low high drift rates. Jou et al. (2016) found the slowest RTs for null lure trials, a finding consistent with race models, but also found slower
RTs for null target trials than for valid trials, a finding consistent with relative evidence models. Additionally, the eyetracking data of Starns et al. (2017) suggest the contribution of a serial comparison process where relative comparisons are only made if neither item is above the criterion. These findings suggest that 2AFC is considerably more complex than yes/no recognition testing.

**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 to control 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.

**Method**

**Participants.** 107 participants from the University of Newcastle participated in exchange for course credit.

**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. 480 words between 3 and 11 letters in length and were 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.
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 that consisted of 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 with the same suit were presented in a row 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 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 the item from the study list. 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 Figure 2.

![Figure 2](image.png)

**Figure 2.** 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.

**Hierarchical Bayesian modeling**

The model was applied to data using hierarchical Bayesian analysis. Although space constraints do not afford a thorough introduction to hierarchical Bayesian analysis (but see Lee, 2011; Lee & Wagenmakers, 2014; Rouder & Lu, 2005; Shiffrin, Lee, Kim, & Wagenmakers, 2008), we would like to highlight three distinct advantages over conventional methods. First, the approach allows for estimates of both group and participant level parameters, which is advantageous because fitting to group data via averaging and fitting individual participant data can produce distortions in cognitive models (Brown &
Heathcote, 2003; Estes & Maddox, 2005; Heathcote, Brown, & Mewhort, 2000). Second, Bayesian methods allow for the quantification of uncertainty in the parameter estimates as posterior distributions. Third, 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 canonically require 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 et al., 2013), a method of posterior sampling that is robust to parameter correlations. With DE-MCMC, on each MCMC iteration a proposal for a vector of parameters is generated by randomly sampling two other parameter vectors from other MCMC chains, taking a scaled difference between those parameter vectors, and then adding that scaled difference to the current parameter vector. This solves the problem of parameter correlation because proposals are generated by sampling from a correlated surface of accepted proposals. We encourage interested readers to consult the Turner et al. (2013) article for a detailed and technical description of this procedure.

For the present experiment, non-informative priors were employed on the memory model parameters while minimally informative priors were used on the parameters specific to the DDM; these prior distributions can be found in Appendix A.
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 25th iteration. The process continued until 1,000 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.

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

**Parameterization**

A complete list of all model parameters for the present experiment can be found in Table. To simplify the model further, we fixed the learning rate $r$ to 1 for the present application (although later fits allow the parameter to vary across conditions varying in the number of presentations). The effect of WF on memory performance was modeled by allowing different values of $\rho$ for each WF class. To improve sampling, an average value $\rho_{\text{avg}}$ along with an effect size parameter $\rho_{\text{effect}}$ were sampled, where $\rho_{\text{effect}}$ is defined as a proportion of $\rho_{\text{avg}}$ and is bounded on a $[0,1]$ interval. $\rho_{LF}$ and $\rho_{HF}$ were defined as follows:

\[
\rho_{LF} = \rho_{\text{avg}} - (\rho_{\text{avg}}\rho_{\text{effect}}) \\
\rho_{HF} = \rho_{\text{avg}} + (\rho_{\text{avg}}\rho_{\text{effect}})
\]

In the immediate testing conditions, $\mu_{ss}$ for study items was fixed to 1 for the first trial. In delayed conditions, the value of $\mu_{ss}$ for study items on trial 1 was estimated as a free parameter $\mu_{ss,\text{delay}}$ to reflect the weaker contextual match after the larger retention interval.
Because manipulations of both list length and retention interval are cross-list manipulations, we allow $z/a_{\text{intercept}}$ and $a_{\text{intercept}}$ to vary across each of these conditions (cf. Ratcliff & McKoon, 2008). These parameters provide an additional confound in the interpretation of list length designs not discussed by Dennis et al. (2008); if it is the case that participants set shallower response boundaries in longer lists, it is possible that observed decrements in longer lists do not reflect worse memory performance but is instead a consequence of how participants set their decision thresholds. In the domain of short-term recognition memory, Donkin and Nosofsky (2012b) found that the best performing model in their model selection was one where only response boundaries changed as list length was increased to capture the slowing of RT with longer lists. Given the relatively short number of trials in the short list, we allow $z/a_{\text{slope}}$ and $a_{\text{slope}}$ to vary across the immediate and delayed conditions but do not allow them to vary across the list length conditions.

Prior work with the DDM has shown that 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 have avoided this assumption for several reasons. First, none of the datasets we used contain a speed emphasis condition. In addition, some of our prior work has found that the DDM with no starting point variability provides a simpler account of recognition memory data than when $s_z$ is estimated as a free parameter (Osth, Bora, et al., 2017).

As mentioned previously, only a single value of $\sigma_{tt}^2$, $\sigma_{ti}^2$, $\gamma$, and $d_c$ were estimated for each participant. In the DDM, a single value of $t_o$ and $s_t$ were estimated for each participant. All model parameters and the experimental factors they vary across can be found in Table 2.
Table 2

Model parameterizations and number of parameters ($N$) for each dataset.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
<td>$0 : \infty$</td>
<td>$r = 1$</td>
<td>R</td>
<td>R</td>
<td>1</td>
</tr>
<tr>
<td>$\mu_{ss}$</td>
<td>$0 : 1$</td>
<td>$\mu_{ss,imm} = 1$, $\mu_{ss,del} \mu_{ss,1} = 1$</td>
<td>$\mu_{ss,1} = 1$</td>
<td>$\mu_{ss,1} = 1$</td>
<td>$\mu_{ss,1} = 1$</td>
</tr>
<tr>
<td>$\mu_{tt}$</td>
<td>$0 : 1$</td>
<td>$\mu_{tt} = 1$</td>
<td>$\mu_{tt} = 1$</td>
<td>$\mu_{tt} = 1$</td>
<td>$\mu_{tt,LD} = 1$, $\mu_{tt,bl} = 1$</td>
</tr>
<tr>
<td>$\sigma^2_{tt}$</td>
<td>$0 : \infty$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>T*</td>
</tr>
<tr>
<td>$\rho$</td>
<td>$0 : \infty$</td>
<td>F</td>
<td>F</td>
<td>$\rho = .04$</td>
<td>$\rho = .04$</td>
</tr>
<tr>
<td>$\sigma^2_{ti}$</td>
<td>$0 : \infty$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>$0 : 1$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\beta$</td>
<td>$0 : \infty$</td>
<td>$\beta = .05$</td>
<td>$\beta = .05$</td>
<td>$\beta = .05$</td>
<td>$\beta = .05$</td>
</tr>
<tr>
<td>$d_c$</td>
<td>$-\infty : \infty$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$z/a_{intercept}$</td>
<td>$0 : 1$</td>
<td>L,D</td>
<td>F,R</td>
<td>1</td>
<td>T</td>
</tr>
<tr>
<td>$z/a_{slope}$</td>
<td>$-\infty : \infty$</td>
<td>D</td>
<td>F,R</td>
<td>1</td>
<td>T</td>
</tr>
<tr>
<td>$a_{intercept}$</td>
<td>$0 : 1$</td>
<td>L,D</td>
<td>F,R</td>
<td>1</td>
<td>T</td>
</tr>
<tr>
<td>$a_{slope}$</td>
<td>$-\infty : \infty$</td>
<td>D</td>
<td>F,R</td>
<td>1</td>
<td>T</td>
</tr>
<tr>
<td>$t_{ER}$</td>
<td>$0 : \infty$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>T</td>
</tr>
<tr>
<td>$s_t$</td>
<td>$0 : \infty$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>T</td>
</tr>
<tr>
<td>$N$</td>
<td>-</td>
<td>21</td>
<td>26</td>
<td>12</td>
<td>19</td>
</tr>
</tbody>
</table>

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. 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. * = For this case, informed priors on $\sigma^2_{tt}$ were only used to constrain the estimate of this parameter in the blank condition. See the text for details.
Posterior predictives

In order to evaluate how well the model matches the data, posterior predictives were generated by generating predictions from the model. 3.33% of posterior samples from each participant’s posterior distributions were selected (1 in every 30th sample) and predictions from the model were simulated. To evaluate how well the model captures the TPE, Figure 3 depicts data and predictions for the short list along with the long list, which was broken up into four blocks of 24 trials. Depicted in the figure are choice probabilities (top) along with correct and error RTs (middle and bottom panels) 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, .5, and .75 quantiles of the RT distribution (the RTs at the 25th, 50th, and 75th percentiles of each participant’s 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). Similar to findings from Criss et al. (2011), the TPE is most evident as a decline in the HR over the test blocks ($\Delta HR = -.08$ in both conditions), although there are slight increases in the FAR as well ($\Delta FAR = .02$ in both conditions). The TPE is accompanied by slower correct RTs through the course of testing (previously reported in Murdock & Anderson, 1975), these are most evident in the later quantiles of the RT distribution. The model captures all of these trends.

In order to evaluate how the model is capturing other aspects of the data, predictions and data for each of the other factors of the experiment, namely word frequency, study-test
Figure 3. Data and posterior predictive distributions for the short list and four blocks of the long list (24 trials in each block) for the immediate (left) and delayed (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 .25, .5, and .75 quantiles of the RT distribution, while the error RTs are summarized using the mean of each participant’s median RT.

delay, and list length, can be seen in Figure 4. One should note that the data and model predictions for the long list condition span the entire 96 test trials of the long list, meaning they consider the effect of the TPE in addition to interference from the list length
manipulation itself. The figure restricts consideration to each factor individually to pool a greater number of trials, which allowed for reliable averages of the .1, .5, and .9 quantiles of the RT distributions. One can see that the model is capturing the patterns in both the choice probabilities and RTs for each manipulation in the experiment, suggesting that the model is not only capable of capturing the TPE but several other manipulations that are constraints on models of recognition memory.

Parameter estimates

The posterior predictives revealed that the model is able to reproduce the key trends in the data. In order to assess which parameters are most responsible for the TPE, we generated a predicted discriminability decrement, \( d'_{\text{change}} \) 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 \( d'_{96} \). This was accomplished by producing 50,000 simulations for a target and a lure in the first and last trial in each condition. This was done for 10% of posterior samples for each participant.

Correlations between \( d'_{\text{change}} \) and the three major culprits of the TPE, namely the context drift parameter \( \gamma \), the item mismatch variability parameter \( \sigma^2_{ti} \) which governs the total degree of item noise, and the total change in response boundaries over the test sequence \( a_{\text{change}} \) (where \( a_{\text{change}} \) is \( a_{\text{slope}} \) multiplied by the number of test trials in the long list). These correlations collapsed over both posterior samples and participants. Due to the large degree of variability in \( d'_{\text{change}} \) over parameters, we reduced consideration to each participant’s 95% highest density interval (HDI), although analyses with the full posterior distribution produced largely similar results. Scatterplots along with 2D kernel density estimates of each of these comparisons for the immediate and delayed conditions can be seen in Figure 5.

Both the immediate and delayed conditions display largely similar results. \( \gamma \), which is inversely related to the overall degree of context drift and the magnitude of the TPE,
Figure 4. Data and posterior predictive distributions for the manipulations of word frequency (WF, top), study-test delay (middle), and list length (bottom). Depicted are choice probabilities (left column), correct RTs (middle column), and error RTs (right). RT distributions are summarized using the average of each participant’s .1, .5, and .9 quantiles of the RT distribution.

shows large correlations in each condition \(r \sim .65\). \(\sigma^2_{ti}\), which is positively related to the overall degree of item noise and the magnitude of the TPE, showed mild negative correlations in each condition \(r \sim -.075\). Surprisingly, changes in response boundaries showed substantial correlations in each condition \(r_{immed} = .24, r_{delayed} = .45\). Further inspection of the scatterplots reveals that a non-trivial percentage of posterior samples
Figure 5. Scatterplots, 2D density estimates, and correlation coefficients of the predicted $d'$ decline through the long test list against the parameters $\gamma$, $\sigma^2_{ti}$, and $a_{change}$ for the immediate (top row) and delayed (bottom row) conditions. Darker areas of the colored density depict the areas with the highest posterior density.

(13.1%) show a predicted increase in performance through the course of testing. These posterior samples are almost entirely restricted to cases where participants were increasing their response boundaries through the test sequence ($a_{change} > 0$).

To get a sense of the inter-participant variability in the changes and response bias and boundaries, the median of each participant’s slope 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 Figure 6. Inspection of the figure reveals that participants become more conservative, as evidenced by a decreasing response bias through the test in both the immediate and delayed condition. This was confirmed by an analysis.
of the group mean parameters for each condition, which did not include zero in their 95% HDI (immediate: $M = -.075 [-.099, -.053]$, delayed: $M = -.042 [-.067, -.017]$, 95% HDI in brackets). Changes in response boundary, in contrast, did not show a consistent relationship in each condition. Nonetheless, group means of the change in response boundary were slightly negative ($M = -.099 [-.18, -.006]$) and the 95% HDI did not include zero, while the delayed condition showed a slight increase ($M = .067, [-.025, .166]$) that included zero in its 95% HDI. Thus, the strong correlations between response boundaries and the TPE were likely because there were large variations across participants in their magnitude of $a_{slope}$ as there were not strong tendencies in either direction in the group means.

![Changes in Bias and Decision Boundaries](image)

*Figure 6.* Histograms depicting the magnitude of changes in $z/a$ and $a$ over the test sequence in the long list for both the immediate (top) and delayed (bottom) conditions. Depicted are the median of each participant’s posterior distribution for each respective parameter.
While item noise appears to be less responsible for the TPE than context drift and changes in response boundaries, it remains possible that item noise is responsible for the list length effect. To investigate this, we performed a similar analysis where we compared the predicted list length effect, $d'_{\text{long}} - d'_{\text{short}}$, to the item mismatch variability parameter $\sigma^2_{ti}$ along with the difference in response boundaries between the short and long list conditions, $a_{\text{long}} - a_{\text{short}}$. 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 Figure 7. Somewhat larger correlations were found between $\sigma^2_{ti}$ and the predicted list length effect, with larger correlations in the immediate condition ($r = -0.452$) than the delayed condition ($r = -0.239$). However, similar to the case with the TPE, 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 parameters between the short and long list conditions: neither the immediate ($M = 0.007 [-0.11, 0.13]$) or the delayed ($M = -0.07, [-0.19, 0.06]$) conditions had difference distributions that included zero in their 95% HDIs. Thus, it appears that the large variability across participants in shifting response boundaries across the list length conditions appears to be responsible for the very strong correlations between $a_{\text{long}} - a_{\text{short}}$ and the list length effect.

Model selection

One potential concern with our model is that the comprehensiveness of its components, including context drift, item noise, and changes in response bias and boundaries, may lead to overfitting. To counteract this concern, we fit a number of simpler models to the data and used model selection techniques which quantify the complexity of the model and subtract that from a measure of its ability to fit the data. In this article, we employ the widely applicable information criterion (WAIC: Watanabe, 2010). In WAIC, model complexity is measured by the variability in the likelihood of a data point across posterior samples summed across all data points. We preferred WAIC to DIC.
Figure 7. Scatterplots, 2D density estimates, and correlation coefficients of the predicted list length effect $d'_{long} - d'_{short}$ against $\sigma^2_{ti}$ and $a_{long} - a_{short}$ for the immediate (top row) and delayed (bottom row) conditions. Darker areas of the colored density depict the areas with the highest posterior density.

(Spiegelhalter, Best, Carlin, & van der Linde, 2002), which is used in MCMC sampling software such as WinBUGS and JAGS, because it is much more numerically stable than DIC (see ?, ?). WAIC is an approximation to leave-out-one cross validation, and so chooses a model on the basis of its ability to predict new data. Smaller values of WAIC mean that a model gives better out-of-sample predictions by striking a better balance between goodness-of-fit and simplicity.

We fit a total of three additional models to the data, each of which fixes one or more parameters to eliminate estimation of one of the aforementioned components: a model with $\sigma^2_{ti} = 0$, which completely lacks item noise, a model with $\gamma = 1$, which lacks context drift,
and a model where \( z/a_{slope} \) and \( a_{slope} = 0 \), in which there are no changes in response bias and boundaries across trials. Because WAIC is only meaningful when compared to another model, we calculated \( \Delta \text{WAIC} \) 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, \( \Delta \text{WAIC} \) scores greater than 10 are conventionally considered large. \( \Delta \text{WAIC} \) values for each model for all of the datasets in this article can be seen in Table 3.

<table>
<thead>
<tr>
<th>( \sigma^2_{ti} = 0 )</th>
<th>LL Exp. (2010, E2)</th>
<th>Criss (2014, E2)</th>
<th>Starns (2014, E2)</th>
<th>Annis et al. (2016, E2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \gamma = 1 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( z/a_{slope}, a_{slope} = 0 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: \( \sigma^2_{ti} \): item mismatch variability parameter, \( \gamma \): context drift parameter, \( \Delta \text{WAIC} \): WAIC difference between the restricted model and the full model, with positive values indicating improvement over the full model.

One can see that none of the restricted models outperformed the full model. Despite the weak correlations between item noise and the magnitude of the TPE, the full model still outperforms the restricted \( \sigma^2_{ti} = 0 \) model for Experiment 1. Nonetheless, the \( \Delta \text{WAIC} \) decrement for the \( \sigma^2_{ti} \) was the smallest of the restricted models, and was virtually negligible for the fit to the data from Annis, Dube, and Malmberg (2016). The \( z/a_{slope}, a_{slope} = 0 \) model showed the largest penalty of the models for Experiment 1 and for the data from Starns (2014), suggesting that the variability across participants in the changes in \( z/a \) and \( a \) through testing considerably improves the model’s ability to capture each participant’s
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 the change in performance. This was likely because the list length manipulation produced only a small effect on performance, consistent with previous investigations. The context drift process, in contrast, appeared to be most predictive of the TPE in both conditions. 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 in the direction of the boundary shift in each condition.

Pure Strength Lists: Applying the model to data from Criss (2010, Experiment 2)

Recently, Kiliç et al. (2017) uncovered additional trends that they argued uniquely supported the REM model of the TPE (Criss et al., 2011). Specifically, they found that when pure strength lists were used, HRs declined at a shallower rate for weak items than for strong items. However, when a mixed strength list was employed, HRs declined at roughly the same rate for weak and strong items. Their variant of the REM model employs what is called a differentiation process during testing and predicted both qualitative trends. In differentiation models, repetitions of items accumulate into a single strong trace that responds more strongly to its own cue but responds more weakly to other cues. Functionally, differentiation produces a reduction in item noise 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. What this entails is that in pure strong lists,
differentiation occurs more often for studied targets, reducing the item noise and the overall
TPE. In pure weak lists, differentiation occurs less frequently, producing a steeper TPE.

Interestingly, context drift in combination with the changes in response bias appears
to make the same predictions for both pure and mixed strength lists. To demonstrate this,
we averaged all of the participant parameters from the immediate condition of the previous
dataset for the $\sigma_t^2 = 0$ model (which lacks item noise) and simulated performance for the
experiment. To simulate performance in the weak and strong conditions, the learning rate
$\alpha$ was set to .5 and 2.0 for each respective condition. For the mixed strength case, expected
d and $S$ parameters were generated using $\alpha = 1.25$ (the average of the weak and strong
learning rates) and the likelihood ratio transformation was calculated according to
Equations 17 and 18.

HR and FAR predictions from the model for each test trial in pure and mixed
strength lists can be seen in Figure 8. For the pure strength lists, HR declines at a much
steeper rate for weak ($\Delta HR = -.10$) than for strong items ($\Delta HR = -.04$). For the mixed
strength lists, the HR declines are more similar for weak ($\Delta HR = -.06$) and strong items
($\Delta HR = -.03$). In this simulation, the only critical difference between these two conditions
concerns the nature of the expected strengths in the likelihood ratio transformation. 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
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.

One should note that these predictions only hold when $z/\alpha_{slope}$ is negative; when
Figure 8. 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 $r_{\text{weak}}$ and $r_{\text{strong}}$ which were set to .5 and 2.0, respectively, which are as follows: $z/a_{\text{intercept}} = .59$, $z/a_{\text{slope}} = -0.0008$, $a_{\text{intercept}} = 1.81$, $a_{\text{slope}} = -0.0009$, $d_c = .062$, $\sigma^2_{tt} = .16$, $\rho = .061$, and $\gamma = .997$.

$z/a_{\text{slope}}$ 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 quite common and also replicates the findings of Ratcliff (1978). In addition, pure list manipulations are also complicated by the fact that cross-list manipulations of list strength permit differences in the linear functions on response bias and boundaries across the different conditions. Indeed, investigations of response times using the list strength paradigm have found that participants employ a more lenient starting point in conditions of higher list strength (Criss, 2010; Starns, Ratcliff, & White, 2012; Kiliç & Öztekin, 2014).

To test these predictions from our model, we used data from Criss (2010, Experiment
which used a pure strength manipulation. 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 by using the posteriors from our initial fit as informed priors for the fit to this dataset and for all other fits in this article. 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 (both mean and standard deviation/sample size) of the $\rho$, $\sigma^2_{rt}$, $\sigma^2_{ti}$, and $\gamma$. Density estimation was conducted for a gridsize of 10,000 points on a region 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 $\sigma^2_{ti} = 0$, $\gamma = 0$, 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. The complete list of model parameters for this dataset along with which parameters were constrained by informed priors can be seen in Table 2.

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

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. Because the manipulations of WF and strength were cross-list, different linear functions of response
bias and boundaries were allowed for each condition in the experiment. Additionally, because immediate testing was used in this experiment, the context reinstatement parameter $\mu_{ss}$ was fixed to one. The complete list of parameters can be found in Table 2.

**Posterior predictives**

Figure 9 depicts data and predictions for the four conditions of the experiment. Test lists were blocked into four blocks of 25 trials. Depicted in the figure are 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.

One can see from the figure that the model is providing an excellent account of the data. The TPE is primarily manifested as a decrease in HR over trials, with a steeper decline in the weak conditions ($\Delta$ HF HR = -.21, $\Delta$ LF HR = -.19) relative to the strong conditions ($\Delta$ HF HR = -.05, $\Delta$ LF HR = -.09). The model reproduces this trend, although it underpredicted the magnitude of HR decline in the weak conditions ($M\Delta$ HF HR = -.16, $M\Delta$ LF HR = -.15) and slightly overpredicted the magnitude of decline in the strong HF condition ($M\Delta$ HF HR = -.09, $M\Delta$ LF HR = -.08), but otherwise reproduces the qualitative pattern. FAR showed increases in all conditions that ranged from slight to moderate in magnitude ($\Delta$ HF weak FAR = .04, $\Delta$ LF weak FAR = .05, $\Delta$ HF strong FAR = .07, $\Delta$ LF strong FAR = .01). Additionally, correct RTs increased over the course of testing, while error RT was largely unaffected. Each of these patterns were reproduced by the model.

**Parameter estimates**

Correlations between $d'_{change}$ and the three major culprits of the TPE, namely the context drift parameter $\gamma$, the item mismatch variability parameter $\sigma_t^2$, which governs the total degree of item noise, and the total change in response boundaries over the test
Figure 9. 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. 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.

sequence $a_{\text{change}}$ (where $a_{\text{change}} = 100a_{\text{slope}}$), 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 Figure 10.

Analyses reproduced the key findings from the list length dataset. The context drift parameter $\gamma$ appeared to have the strongest correlation with the TPE, exhibiting
Figure 10. Scatterplots, 2D density estimates, and correlation coefficients of the predicted $d'$ decline through the test list against the parameters $\gamma$, $\sigma^2_{ti}$, and $a_{\text{change}}$ for each condition of the Criss (2010, Experiment 2) dataset. Darker areas of the colored density depict the areas with the highest posterior density.

correlations ranging from .7 to .83 across each condition. $a_{\text{change}}$ showed strong correlations ranging from .38 to .5. The item mismatch variability parameter, $\sigma^2_{ti}$, showed extremely weak correlations, with the largest correlation being -.08 (in the LF strong condition).
These analyses converge with prior analyses suggesting that item noise appears to play only a minor role in producing the TPE. To get a sense of the inter-participant variability in the changes and response bias and boundaries, 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 Figure 11. Inspection of the figure reveals that participants become more conservative in the weak conditions, as measured by negative values of $z/a_{\text{change}}$. For the strong conditions, in contrast, there was not an overall tendency to change bias through the course of the test. This dissociation was reflected in the group mean parameters for each condition. The 95% HDIs for $z/a_{\text{change}}$ for the HF ($M = -.12 [-.17, -.08]$) and LF weak ($M = -.13 [-.17, -.09]$) conditions did not include zero, whereas zero was included in the HDIs for the HF ($M = -.01 [-.07, .05]$) and LF ($M = -.005 [-.05, .05]$) strong conditions. 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 through the test.

Changes in response boundaries ($a_{\text{change}}$) showed similar results as the parameter estimates from the list length dataset. Despite a substantial degree of variability across participants, participants did not show a consistent tendency in their changes in response boundaries through the test. All of the HDIs for the group mean parameters in each condition included zero.

Discussion

In line with the predictions in Figure 8, the model was able to reproduce the greater decline of HR in pure strong lists relative to pure weak lists. In addition, analyses of the posterior predictives of the TPE demonstrated that context drift and changes in response boundaries were primarily responsible for the TPE. To test the predictions for mixed strength study lists, we employed data from Starns (2014, Experiment 2).
Mixed Strength Lists: Applying the model to data from Starns (2014, Experiment 2)

As mentioned previously, a prediction of our model is that the HR decline should be similar for weak and strong items in mixed lists of strong and weak items. To test this prediction, we applied our 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. Because there were no cross-list manipulations, only a single linear function for response bias and boundaries was employed for each participant. 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.

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

**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 were used relatively uninformative priors. In addition, because words appeared to be of medium frequency in this dataset, we fixed $\rho$ to .04, which was similar to the group mean estimate of $\rho_{avg}$ from the previous dataset ($M = .038$). 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 used in this experiment and thus the context reinstatement parameter $\mu_{ss}$ was fixed to one. The complete list of parameters can be found in Table 2.

**Posterior predictives**

Figure 12 depicts data and predictions for the experiment. Test lists were blocked into four blocks of 25 trials. Depicted in the figure are 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 the data. In contrary to the pure strength manipulations in the Criss dataset, this dataset shows very similar declines in HR across the two strength conditions ($\Delta$ weak HR = -.09, $\Delta$ strong HR = -.07). The model produces virtually identical declines across the
Figure 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.

weak and strong conditions ($M \Delta HR = .07$ in both conditions). In addition, there was a moderate increase in the FAR ($\Delta FAR = .05$), increasing correct RTs through the course of testing, while error RTs were not strongly affected by the course of testing. The model reproduces each of these qualitative trends.
Parameter estimates

Correlations between $d'_{\text{change}}$ and the three major culprits of the TPE, namely the context drift parameter $\gamma$, the item mismatch variability parameter $\sigma^2_{t_i}$ which governs the total degree of item noise, and the total change in response boundaries over the test sequence $\Delta_{\text{change}}$ (where $\Delta_{\text{change}} = 100a_{\text{slope}}$), 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 Figure 10.

![Figure 10](image_url)

**Figure 10.** Scatterplots, 2D density estimates, and correlation coefficients of the predicted $d'$ decline through the test list against the parameters $\gamma$, $\sigma^2_{t_i}$, and $\Delta_{\text{change}}$ for the weak (top) and strong (bottom) conditions of the Starns (2014, Experiment 2) dataset. Darker areas of the colored density depict the areas with the highest posterior density.

Similar to previous analyses, the biggest predictor of the TPE was the context drift
parameter $\gamma$, producing correlations $\sim .7$ in each condition. Changes in response boundaries produced moderate correlations ($r \sim .325$) while the item mismatch variability parameter $\sigma_{ti}^2$ produced quite small correlations. These correlations were also noticeably in the wrong direction: increasing item noise was associated with less of a decline in performance in across trials.

Histograms of the changes in bias and response boundary parameters can be seen in Figure 14. There was not a strong tendency to change either bias or response boundaries over trials, as the 95% HDIs on the group means included zero for both $z/a_{\text{change}}$ ($M = -.01$, [-.05, .03]) and $a_{\text{change}}$ ($M = .04$, [-.04, .11]). 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. Similar to previous results, this suggests that the changes in response boundaries accounts for individual variation in the magnitude of the TPE but does not produce the TPE more generally.

Discussion

In line with the predictions in Figure 8, 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. In addition, analyses of the posterior predictives of the TPE demonstrated that context drift and changes in response boundaries were the biggest predictors of the TPE.

The effect of task switching on the TPE: Data from Annis, Dube, and Malmberg (2016, Experiment 2)

A remaining puzzle for models of the TPE was presented 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: one condition where participants had a long interstimulus interval (ISI) between trials, which we will refer to as the 'blank' condition. In a separate 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 decrements to the blank condition, despite the fact that the effective number of test stimuli is twice as large in such conditions. Annis et al. (2013) noted that this result is seemingly contrary to contextual drift accounts of the TPE, which suggests that such a result may be challenging for our model.

However, one consequence of task switching manipulations 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). A discovery with our model is that increases in response boundaries actually reduce the TPE. Using the same parameters as in Figure 8, we generated model predictions for several different values of the response boundary parameter $a$ with a sequence of 160 test trials. For the shallowest response boundary, the decline in HR was steep over the test list ($\Delta HR = -.16$). For the most cautious response boundary, the decline in HR was considerably shallower ($\Delta HR = -.07$). What this suggests is that the task-switching manipulation may have incidentally introduced a factor that mitigated against the greater contextual change and increasing item noise of the LD trials. While the data from Annis et al. (2013) utilized the 2AFC paradigm, we applied the model to an analogous experiment utilizing the yes/no format was conducted by Annis et al. (2016, Experiment 2).

To ensure that the higher response boundaries in the LD condition were not the consequence of overfitting, we additionally fit the standard DDM to the data with drift rate distributions estimated as free parameters. We compared two models, a model where the response boundary $a$ varied across the blank and LD conditions (the $a \sim T$ model) and a model where $a$ was fixed across both conditions. Model selection exhibited a very large preference for the $a \sim T$ model, which also demonstrated substantially higher response boundaries in the LD condition relative to the blank condition. Details involving these models and the model fitting procedure can be found in Appendix B.

In fitting this dataset, we assumed that LD trials contribute both additional item noise and cause context drift, meaning that the influence of these factors should be twice as large in the LD condition relative to the blank condition. In actuality, it may be possible that the task context of the LD trials is radically different from the recognition test trials, such that they do not cause item interference or alter the context cue used for recognition testing. Nonetheless, at least one prominent model of episodic memory, namely the context maintenance and retrieval model (CMR: Polyn et al., 2009), assumes that task-switching causes a change in the episodic context in a similar manner as new items from within the
Figure 15. Simulated predictions of HR and FAR from the model with three different intercepts for the response boundary parameter $a$.

same task. Similarly, investigations into the effect of testing on context change have found that semantic retrieval induces similar degrees of context change as episodic retrieval (Divis & Benjamin, 2014; Jang & Huber, 2008). For this reason, and to provide the strongest challenge for our model, we fit the model with the assumption that the LD trials produce the same functional effect as the recognition test trials.

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 seconds were removed. Three participants were excluded for having over 20% of their RTs excluded under such criteria. Subsequently, we removed all RTs that were larger than 3 standard deviations above a participant’s mean RT. The first test trial from each block was removed due to higher RT ($M = 3.25$ seconds) than the grand mean ($M = 1.40$ seconds). This resulted in the exclusion of 3.5% of all responses.
Parameterization

As mentioned previously, 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, $\mu_{tt}$, 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). In addition, we assumed that there is higher encoding variability in the LD condition relative to the blank condition. We ensured that $\sigma^2_{tt}$ was higher in the LD condition by sampling $\sigma^2_{tt,\text{increment}}$ and defining $\sigma^2_{tt,\text{LD}} = \sigma^2_{tt,\text{blank}} + \sigma^2_{tt,\text{increment}}$. Because the blank condition was analogous to conditions from the prior fits, $\sigma^2_{tt,\text{blank}}$ was constrained using inforemd priors from the fit to the Starns et al. (2014) dataset while $\sigma^2_{tt,\text{increment}}$ was estimated using relatively uninformative priors. No other memory related parameters were allowed to vary in the LD condition.

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 $\mu_{ss}$ was fixed to one. 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 2.
Posterior predictives

Figure 16 depicts data and predictions for the experiment. Test lists were blocked into four blocks of 40 trials. Depicted in the figure are 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 LD condition ($\Delta HR = -.14$) relative to the blanks condition where there is no intervening task between recognition trials ($\Delta HR = -.16$). This is surprising when one considers that when one additionally 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 ($M \Delta HR LD = -.07$, $M \Delta HR$ blank = -.11) and appears to be overpredicting the level of performance in the blanks condition. In contrast to previous datasets, the FAR decreases in the blank condition ($\Delta FAR = -.04$), 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 increases in the FAR in each condition ($\Delta FAR$ blank = .01, $\Delta FAR$ LD = .06).

In addition, there were large increases in RT in the LD condition relative to the blank condition which shifted the entire RT distribution, which the model is able to capture. Aside from the aforementioned deviations between the model and the data, the model appears to be providing a very good account of the data.

Parameter estimates

Correlations between $d'_{\text{change}}$ and the three major culprits of the TPE, namely the context drift parameter $\gamma$, the item mismatch variability parameter $\sigma^2_{ti}$ which governs the total degree of item noise, and the total change in response boundaries over the test
Figure 16. 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 for weak items and the median RT for the strong condition and lures. One should note that the test blocks in the LD condition refer to the recognition test trials.

sequence $a_{change}$ (where $a_{change} = n a_{slope}$, where $n$ is the number of test trials), were calculated using the same procedure as before, with the exception that we additionally included the $a_{intercept}$ parameter based on the demonstration in Figure 15. Scatterplots
along with 2D kernel density estimates of each of these comparisons for each experimental condition can be seen in Figure 17.

Figure 17. Scatterplots, 2D density estimates, and correlation coefficients of the predicted $d'$ decline through the test list against the parameters $\gamma$, $\sigma_{ti}^2$, $a_{\text{change}}$, and $a_{\text{intercept}}$ for the Blank and LD conditions of Annis et al. (2016) Experiment 2. Darker areas of the colored density depict the areas with the highest posterior density.

Consistent with all previous analyses, the biggest predictor of the TPE was the context drift parameter $\gamma$, producing correlations of .78 and .57 in the blank and LD conditions, respectively. The item mismatch variability parameter $\sigma_{ti}^2$, in contrast, produced only weak correlations with the predicted TPE ($r \sim -.02$). In contrast to prior analyses, changes in response boundaries produced only moderate correlations with the TPE, with correlations of around .1 for each condition. The intercept of the response boundary function was also negatively correlated in the blank condition, but exhibited the correctly
predicted positive relationship in the LD condition, suggesting that higher response boundaries are decreasing the TPE. Nonetheless, it was perplexing that the correlations even in this condition were so weak in magnitude. It’s possible that the correlations were strongly affected by negative inter-parameter correlations, as the correlation between $a_{slope}$ and $a_{intercept}$ were -.13 in the blank condition and as strong as -.46 in the LD condition.

Histograms of the changes in bias and response boundary along with the intercept of the response boundary parameters can be seen in Figure 18. 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 ($M_{blank} = -.13$, [-.16, -.10], $M_{LD} = -.08$, [-.12, -.05]) and response boundaries ($M_{blank} = -.24$, [-.33, -.15], $M_{LD} = -.34$, [-.50, -.18]) did not include zero in their posterior distributions. Participants also exhibited considerably higher intercepts on the response boundary function in the LD condition ($M = 2.33$, [2.19, 2.43]) than the blank condition ($M = 1.82$, [1.72, 1.92]).

**Discussion**

In line with the predictions in Figure 15, the model, which relied on considerably higher response boundaries to address the long RTs induced by the task-switching component of the LD condition, was able to produce nearly equivalent declines in performance in the LD and blank conditions. This occurred despite the fact that we made the strong assumption that items on the LD trials contribute equivalent contributions to both the item noise and context drift components of the model, effectively doubling the number of test trials in the LD condition relative to the blank condition. This provides further support for the contention that RT data places important constraint on models of memory, as one might misleading conclude from the choice proportions alone that additional assumptions about either the item noise or context drift components of memory models would be required to produce the near identical declines in performance in the two experimental conditions.
Figure 18. Histograms depicting the magnitude of changes in $z/a$ and $a$ over the test sequence in the fit to Starns (2014, Experiment 2).

One should note that other assumptions about the nature of the task-switching manipulation are possible. For instance, it might be that the context representation of the lexical decision task is radically different from the recognition test context to the point where items stored in lexical decision produce no item noise and do not change the contents of the recognition context. We implemented such a model by only assuming that the recognition test trials increment item noise and cause context drift. Nonetheless, model selection exhibited a slight preference for our model that assumes that LD trials contribute both item noise and context drift, with a WAIC difference of around 3 points. For this reason, we do not elaborate on this alternative model further.
Ratios of Target to Lure Variability

As mentioned previously, 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 self match variability parameter $\sigma_{tt}^2$ was higher than the item mismatch variability parameter $\sigma_{ti}^2$. This can be seen in the expression for target variance (Equation 9) which differs from the equation for lure variance (Equation 10) because of the self match term in Equation 9 which contains $\sigma_{tt}^2$. While we lack any datasets that show a bias manipulation that 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 estimated higher drift rate variability for targets than for lures (Osth, Dennis, & Heathcote, 2017; Osth, Bora, et al., 2017; Starns & Ratcliff, 2014; Starns, Ratcliff, & McKoon, 2012; Starns, 2014).

We constructed estimates of the target-to-lure variability $S$ using Equations 9 and 10 with our estimated group mean parameters in each dataset. The results are depicted in Figure 19. Because $S$ changes over the course of testing, we are depicting the estimates predicted for the first trial only. One can see from the figure that estimates of $S$ are higher than 1 in all conditions. 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 $\sigma_{tt}^2$ in each of these conditions. An exception to the rule is the LD condition of the Annis et al. (2016) dataset, where a larger value of $\sigma_{tt}^2$ was estimated in the LD condition to capture the poorer performance as a consequence of the task-switching. Equations 9 and 10 also reveal how the model produces higher variability in conditions of higher performance: consider if one were to manipulate the context mismatch variability parameter $\rho$ to capture different levels of word frequency. This has the effect of increasing the context noise variability, which affects increases the
variability for targets and lures by the same amount, decreasing the ratio.

![Figure 19. Violin plots showing the posterior distribution of target-to-lure variability $S$ for each condition constructed from the group mean parameters.](image)

On the surface, this stands in contrast to investigations using the DDM which have been unable to 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 Figure 19 reveals that the uncertainty in the estimates of $S$ are quite large, which has also been demonstrated with the standard DDM (Osth, Bora, et al., 2017; Osth, Dennis, & Heathcote, 2017). Second, while models that freely estimate $S$ do not estimate differences in $S$ across conditions, models such as ours which impose a relationship between $S$ and the mean drift rate may still be quite adequate for the data. Finally, investigations of confidence and response time with the response time confidence model (RTCON: Ratcliff & Starns, 2009) demonstrated an increasing value of $S$ in conditions of higher performance, similar to our model.
General Discussion

We have introduced a new model of recognition memory based on the Osth and Dennis (2015) model of recognition memory, which uses a matrix of item-context associations where the parameters specify the match distributions between items and contexts, along with a back-end diffusion process to form a complete model of memory retrieval and decision making. The model is ideal for investigating the causes of the detrimental effect of testing on recognition memory, as there are three primary culprits behind the effect, all of which are specified within the model. In particular, the memory retrieval components from the Osth and Dennis (2015) model allow for the estimation of both item noise and changes in the retrieval context, while the diffusion decision model (DDM) allows for the estimation of changes in response bias and boundaries due to its ability to separate the contributions of memory strength and decision factors based on their differential effects on RT distributions. We have applied the 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 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.

A consistent story 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 at its largest, a small role in the decline in performance. Changes in response boundaries additionally showed strong correlations with the change in response boundaries, but in most of the analyses there was no consistent across participants in the direction of their boundary shift. Instead, it appeared there were
large individual differences in shifting response boundaries, with some participants increasing their response boundaries and some participants 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. However, this dataset differed from the others in that the size of the decrease in response boundaries showed only a small correlation with the size of the test position effect (TPE). Thus, it appears that changes in response boundaries play a significant role in moderating the size of the TPE, but this factor does not appear to be sufficient to drive the effect on its own.

Criss et al. (2011) criticized context change as an ad hoc explanation for explaining the TPE. To the contrary, a number of 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). Similarly, a number of free recall models, including the model of Davelaar et al. (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 causing context change. Thus, there appears to be strong independent evidence for the assumption that retrieval causes context change. Similarly, the mechanism of context drift has a lot of explanatory power outside of the effects of 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), 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 contribution in explaining the TPE.

To implement item noise in our model, we adopted an assumption where test items that are added to the contents of memory are maximally similar to the current state of context. This assumption is most assuredly wrong, as it would predict that lures learned during the test list are more similar to the current state of context than studied items, implying the FAR to such lures should be higher than the HR. Investigations with repeated lures have not confirmed such a prediction. Ratcliff and Hockley (1980) evaluated performance for lures presented a second time as a function of their lag from their initial presentations. When the data were conditioned on correct performance on the first presentation, lures that were repeated immediately had very fast RTs and no change in their initial responses. As lag increased from one to ten presentations, participants showed a higher likelihood of false alarming to the second presentation, a tendency which decreased and subsequently asymptoted. We propose here that is more likely that lures are instead bound to a representation of the test context, which may bear some intermediate representation to the study list context, implying that their addition to the contents of memory produces a relatively small contribution to item noise.

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 every case. In addition, our analysis of the list length effect in Experiment 1 showed that item noise played a larger role in explaining the list length effect than its contribution to the TPE. This stands in contrast to the work of Osth and Dennis (2015), who found that item noise played an extremely small role in explaining recognition memory performance for words. Our work deviates from this work in a few important ways. First, we extended the model to response times, which has been shown to introduce considerable constraint on cognitive models, in addition to modeling the 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.

**Changes in Bias and Response Boundaries**

One surprising result of our analyses was the consistently large individual variation in participants’ changes in response bias and boundaries. While the response boundaries 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? Some insight could potentially be gained from the self regulating accumulator model (Lee & Dry, 2006; Vickers, 1979; Vickers & Lee, 1998), an accumulator model which determines choice, confidence, and RT from a race between two accumulators, where confidence is determined by the difference in evidence between the two accumulators. In the self regulating accumulator model, changes in the response thresholds 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.

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. Unfortunately, one of the difficulties in applying self regulation is its reliance on the balance-of-evidence mechanism to determine confidence. To date such models have performed relatively poorly in fitting RT distributions (Van Zandt,
2000). 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. While it would 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 the RTCON models are already quite complex and use many more parameters than the relatively simple balance-of-evidence models.

**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 memory evidence that feeds into a back-end decision model to generate predictions of choice and response time. 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' criterion². 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 that cannot be addressed with a simple change in the

²The 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.
speed-accuracy threshold. A compelling demonstration of this was 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. While 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.

However, aside from this finding, there are few other findings in the literature on single item recognition memory that require a dynamic approach such as the class of feature sampling models. 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 is characterized by 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 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; Kılıç & 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) conducted a number of experiments to evaluate whether recognition memory requires the contribution of a recollection process late in retrieval. They manipulated a number of variables such as list length and the number of presentations and varied whether participants were forced to respond slowly or quickly. Slow responses performed better on all variables, but there were no interactions with list length or presentation time, leading Gillund and Shiffrin to reject the notion that these variables required a unique contribution late in retrieval.

An interesting recent result that might suggest the necessity of a non-stationary process is the finding that drift rates in recognition memory are reduced in conditions of speed emphasis relative to accuracy emphasis (Rae, Heathcote, Donkin, Averell, & Brown, 2014; Starns, Ratcliff, & McKoon, 2012). Such a finding could be indicative of a non-stationary process whereby memory strength increases through the course of retrieval, potentially due to the contribution of feature sampling producing a stronger cue late in retrieval relative to early retrieval. Nonetheless, a remaining possibility is that participants in speed emphasis conditions initiate retrieval with deficient cues relative to accuracy emphasis conditions. This can be addressed in our model by varying the $\mu_{tt}$ parameter, which governs cue-to-target strength, across speed and accuracy emphasis conditions. Future work may be needed to distinguish the deficient cuing account from a non-stationary process.

**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. We applied the 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 necessity of considering the dynamics of decision making in recognition memory.
Appendix A

Prior Distributions on Model Parameters

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

\[
\begin{align*}
    z/a_{\text{intercept}} & \sim TN(M_{z\text{intercept}}, \varsigma_{z\text{intercept}}, 0, 1) \\
    z/a_{\text{slope}} & \sim \text{Normal}(M_{z\text{slope}}, \varsigma_{z\text{slope}}) \\
    a_{\text{intercept}} & \sim TN(M_{a\text{intercept}}, \varsigma_{a\text{intercept}}, 0, \infty) \\
    a_{\text{slope}} & \sim \text{Normal}(M_{a\text{slope}}, \varsigma_{a\text{slope}}) \\
    t_{\text{er}} & \sim TN(M_{t\text{er}}, \varsigma_{t\text{er}}, 0, \infty) \\
    s_{t} & \sim TN(M_{s\text{t}}, \varsigma_{s\text{t}}, 0, \infty) \\
    d_{c} & \sim \text{Normal}(M_{d\text{c}}, \varsigma_{d\text{c}})
\end{align*}
\]

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 $a$, $t_{\text{er}}$, and $s_{t}$ are bounded below at 0 but unbounded on the right. The slope parameters $z/a_{\text{slope}}$ and $a_{\text{slope}}$ 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:

\[
\begin{align*}
    \log(\sigma^2_{ti}) & \sim \text{Normal}(M_{\sigma^2_{ti}}, \varsigma_{\sigma^2_{ti}}) \\
    \log(\sigma^2_{tt}) & \sim \text{Normal}(M_{\sigma^2_{tt}}, \varsigma_{\sigma^2_{tt}}) \\
    \log(\rho) & \sim \text{Normal}(M_{\rho}, \varsigma_{\rho}) \\
    \log(r) & \sim \text{Normal}(M_{r}, \varsigma_{r})
\end{align*}
\]
An exception is the $\rho_{effect}$ parameter, where $\rho_{HF} = \rho + (\rho_{effect}\rho)$ and $\rho_{LF} = \rho - (\rho_{effect}\rho)$. Because $\rho_{effect}$ defines the effect size of the word frequency effect as a proportion of $\rho$, it is bounded on the (0,1) interval and is sampled as follows:

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

(37)

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 $\mu$ and sample size $v$:

$$\alpha = \mu v$$

(38)

$$\beta = (1 - \mu)v$$

(39)

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

The following parameters were sampled from reparameterized beta distributions:

$$\gamma \sim rBeta(\mu_{\gamma}, v_{\gamma})$$

(40)

$$\mu_{ss} \sim rBeta(\mu_{\mu_{ss}}, v_{\mu_{ss}})$$

(41)

$$\mu_{tt} \sim rBeta(\mu_{\mu_{tt}}, v_{\mu_{tt}})$$

(42)

For the group level mean ($M$) parameters of the DDM, we used mildly informative priors, several of which were employed by Osth, Dennis, and Heathcote (2017):
\[ M_{\text{ter}} \sim TN(.5,.5,0,\infty) \] (43)

\[ M_{st} \sim TN(.25,.25,0,\infty) \] (44)

\[ M_{z/\text{intercept}} \sim TN(.5,.5,0,1) \] (45)

\[ M_{\text{intercept}} \sim TN(2,2,0,\infty) \] (46)

\[ M_{z/\text{aslope,aslope}} \sim Normal(0,.05) \] (47)

\[ M_{dc} \sim Normal(0,1) \] (48)

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

\[ M_{\sigma t,\sigma t,\rho,\tau} \sim Normal(0,100) \] (49)

\[ M_{\gamma,\mu s,\mu t} \sim rBeta(.5,2) \] (50)

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 \( M_{\text{effect}} \) we used:

\[ M_{\text{effect}} \sim TN(.5,.5,0,1) \] (51)

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

\[ \varsigma_{\text{intercept,dc}} \sim Gamma(1,1) \] (52)

\[ \varsigma_{z,\text{st},\text{t0,aslope,aslope}} \sim Gamma(1,3) \] (53)

For the \( \varsigma \) parameters of the Osth and Dennis (2015) model, we used less informative priors on several of the parameters:
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. See the main text for details.

\[ \varsigma_{\text{att}, \rho, \tau} \sim \text{Gamma}(0.1, 0.1) \]  
\( v_{\mu ss, \mu tt, \gamma} \sim \text{Gamma}(0.1, 1.0) \)  
\[ \varsigma_{\text{effect}} \sim \text{Gamma}(1, 1) \]


Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press. (Unpublished lecture notes, University of California,


Appendix B

DDM Modeling of the Annis et al. (2016, Experiment 2) Data

We employed a measurement model approach to the data from Annis et al. (2016) to evaluate whether response boundaries were affected by the task-switching manipulation. We accomplished this end by applying the standard DDM with free parameters governing the drift rate distributions. To simplify this approach, we avoided the issue of how performance changes across test trials. For this reason, we employed fix values of \( v \), \( s_v \) (the standard deviation of the drift rate distribution), \( z/a \), and \( a \) that did not change across trials. In addition, we did not employ starting point variability in the model. The drift rate parameters were sampled from the following group level distributions:

\[
\begin{align*}
v_{\text{target}} & \sim \text{Normal}(M_{v,\text{target}}, \varsigma_{v,\text{target}}) \\
s_v & \sim TN(M_{sv}, \varsigma_{sv}, 0, \infty)
\end{align*}
\]

We used relatively uninformative priors on the group mean parameters \( M \):

\[
\begin{align*}
M_{v,\text{target}} & \sim \text{Normal}(2, 2) \\
M_{v,\text{liare}} & \sim \text{Normal}(-2, 2) \\
M_{sv} & \sim TN(1, 1, 0, \infty)
\end{align*}
\]

along with the group standard deviation parameters \( \varsigma \):

\[
\varsigma_{v,\text{target}, v,\text{liare}, sv} \sim TN(1, 1, 0, \infty)
\]

We contrasted two models: a model where \( a \) varies across the task switching manipulation (the \( a-T \) model) along with a model where \( a \) is constrained to be the same
across both the blank and LD conditions (the fixed $a$ model). In both models, we allowed $t_{ER}$, $s_t$, $z/a$, $v_{target}$, and $v_{lure}$ to vary across the blank and LD conditions. The $sv$ parameter was allowed to vary across targets and lures to be consistent with previous evidence suggesting that targets have larger values of $sv$ (Osth, Dennis, & Heathcote, 2017; Starns & Ratcliff, 2014; Starns, Ratcliff, & McKoon, 2012) but was not allowed to vary across the two task-switching conditions. The $a^\sim T$ model has 14 parameters while the fixed $a$ model has 13 parameters.

We employed the same exclusion criteria for these data as the models in the text. For both models, the number of chains was set to be equal to three times the number of parameters. Each model was run with a burn-in period of 5,000 samples. Subsequently, one sample was collected for every 25 iterations until a total of 1,000 MCMC samples was collected for each chain. Both models showed strong convergence with a GR statistic that was close to 1.0 for all parameters.

Despite the extra complexity of the $a^\sim T$ model, it was considerably preferred in model selection, with a $\Delta$ WAIC value of 481.57, which is quite large in magnitude. Analysis of the group mean parameters of the $a^\sim T$ model showed substantial differences in the estimate of $a$ across the blank ($M = 1.84, [1.74, 1.95]$) and LD ($M = 2.40, [2.29, 2.51]$), with no overlap in their 95% HDIs.